Chapter 1

Introduction

BIOLOGICAL systems can be considered as complex systems, where complexity is not defined as merely complicated but in terms of the overall behaviour being unpredictable from understanding the component parts. This chapter defines what is meant by the term “complex systems”, how the biological systems studied in this thesis are complex systems, and how complex systems analysis provides useful techniques to further human knowledge of these biological systems.

1.1 Introduction

Complex systems are ones in which the system shows emergent, nonlinear behaviour—they are dynamic open systems in which the behaviour of the whole is greater than the sum of the parts. This emergence may be weak emergence, in which one could theoretically reduce the behaviour into the sum of the parts, but this may be too difficult and more suitable would be a simpler rule for the overall behaviour. It may also be strong emergence, in which the behaviour of the whole is irreducible into that of the behaviour of the component parts considered in isolation or in other combinations. To quote from Anderson (1972),

> *...it seems to me that one may array the sciences roughly linearly in a hierarchy according to the idea: The elementary entities of science X obey the laws of science Y.... But this hierarchy does not imply that science X is "just applied Y." At each stage entirely new laws, concepts and generalizations are necessary, requiring inspiration and creativity to just as great a degree as in the previous one.*

Complex adaptive systems are complex systems in which adaptation occurs—that is, they change in response to their environment (including other agents). This can occur at one or more levels of a complex system. For example, this may take the form of selection, in which agents, or groups of agents, pass on their traits with a higher frequency than agents that are less fit according to some measure. Agents can also learn, given feedback on their performance, and hence modify their behaviour.

Biological systems consist of tiny parts (for example nucleic acids, proteins, sugars, and other organic molecules) that through complex interactions form cells. These cells in turn aggregate to form other complex systems such as humans, and these in turn form larger complex systems, communities, in a heirarchy of complex systems interacting at the boundaries of the levels. How then to deal with this complexity? There are a number of different and interrelated approaches labelled with the overarching title of "complexity science" including statistical mechanics, agent-based modelling, evolutionary programming, cellular automata and nonlinear techniques. These can be considered either as ways of modelling behaviour in complex systems or as approaches to describing the behaviour using various statistics.

My thesis tackles the following complex biological systems:

1. The evolution and analysis of DNA sequences

2. Mutations in DNA sequences
3. Viruses and memes
4. *Drosophila* (fruit fly)
5. The gene network around the *p53* gene, and how this relates to the development of cancer
6. Modelling of the development of tumors and cancer
7. The human brain during sleep
8. Analysis of metabolites from various biological systems under different conditions (Metabolomics)

Topics 2, 3, and 6 all involve computer simulation and mathematical modelling of biological systems, exploring the overall behaviour of the system from the rules governing the interacting parts. Topics 1 through 8 all involve statistical analysis of data from real biological systems. The remainder of this chapter summarises each of the main chapters of my thesis.

1.2 Analysis of DNA sequences

There are a number of tools provided by complex systems analysis, in particular those based around (or related to) correlations and information. These two methods are shown to be interrelated. Also considered are some other signal processing techniques, since they offer us information on a complex system: not just the proteins encoded as genes in DNA, but other sequences of biological function (such as binding sites).

1.3 Mutations in DNA sequences

Mutations arise in DNA sequences through a number of physical and chemical processes. These mutations play a critical role in such complex systems as population genetics of viruses as they spread through a population, in gene networks such as *Drosophila* (fruit fly) and cancer. Chapter Three details this research into both the evolution and analysis of mutations.

1.4 Viruses and memes

Viruses consist of short genetic sequences, in either DNA or RNA, which are then encapsulated into a protein “shell”. Those with RNA sequences are known as ”retroviruses” as first the genetic sequence has to be translated back into DNA before it can be incorporated into the DNA sequence of the host cell. The virus then instructs the cell to make more copies of the virus, unless it remains dormant as some viruses do. Viruses that affect bacteria are known as bacteriophages, and operate in a similar fashion, hijacking the normal host cell processes in order to replicate. Viruses in and of themselves, while certainly complicated systems, are not typically classified as complex systems. There are two well-known ways in which viruses act within complex systems—as part of the host cell’s genetic regulatory network, and in terms of transmission on a social network. Both gene networks (Kauffman 1993) and social networks (Clauset and Moore 2003) are well studied complex systems, exhibiting emergent behaviour.

1.5 *Drosophila*

Drosophila melanogaster, commonly known as fruit fly, is an organism with a long history of very detailed genetic study. This facilitates exploration of the gene network in *Drosophila* larvae that sets up a pattern of stripes, which later develop into segments in the larval stage and then various body parts in the adult fruit fly. A cellular automaton was used to explore this gene network to gain new insights into the gene interactions and the robustness of this system.

1.6 The *p53* gene

As a step between the gene networks in Drosophila and modelling of the growth of cancer, it is important to look at a very detailed level of a critical cancer gene, *p53*, and its associated network. This topic also links back to the chapter on mutations in DNA sequences, since *p53* plays a role in the detection and repair of mutations in DNA sequences. This gene is also involved in a number of other critical pathways in the cellular gene network, including control of the cell cycle and programmed cell death, or apoptosis.

1.7 Cancer

Cancer is a complex system with a very large number of interacting parts. This includes not only the gene networks, but also groups of individual cells within the cancer playing heterogeneous roles in the overall cancer pathology. Chapter Seven presents a model of the growth of cancer, focussed at a general level on genetic changes that occur. The work on *p53*, while worthwhile, offers little in the way of quantitative results, and it is only possible to look at this in depth because the gene network is so well studied. There are many other gene networks involved in cancer that aren't as well studied. The complexity of cancer also arises from the emergent behaviour of the interacting parts, so this is another reason to use complex systems approaches.

1.8 The human brain during sleep

The human brain consists of large numbers of neurons (around 10^{11}), wired with a staggering 10^{14} synapses (connections), and is one of the best examples of a complex system. The emergence of consciousness is no less than amazing. The work in Chapter Eight, analysing the brain during sleep, was carried out with researchers from a number of universities and the Adelaide Women's and Children's Hospital's sleep unit. Tools from complex systems research were applied, and showed that significant nonlinear *and* linear behaviours are present in the brain during sleep. It is also shown how computers can learn to spot overall trends from this extremely complex system.

1.9 Metabolomics

Rather than worry about individual details of gene expression, or even cellular interactions and (higher still) organs of the body, one can analyse the system as a whole through analysis of metabolic output of both groups of cells and organisms as a whole. Metabolomics is the study of metabolic output of biological systems to analyse their inner workings. In the same way that smoke in the exhaust could mean the oil needs to be changed or that the steering wheel shaking might mean the wheels need aligning, so too outputs of a biological system can give information about its innards. The system can be a cell or an organism, and the innards need not be understood in detail. That this is an important complex systems approach is detailed by Bino *et al.* (2004) who write, *For a holistic understanding of the biological behavior of a complex system, it is essential to follow, as unambiguously as possible, the response of an organism to a conditional perturbation at the transcriptome, proteome and metabolome levels.*

Chapter 2

DNA analysis

A Number of complex systems signal processing and statistical methods can be used in analysing DNA sequences. These techniques can be used in a number of ways, from finding genes in DNA to determining phylogenetic trees that show relationships between living organisms. Signal processing methods such as spectrograms provide useful new tools in the area of genomic information science. In particular, fractal analysis of DNA "signals" provides a new way of classifying organisms.

2.1 Introduction

The Human Genome Project (international 2001), together with a number of other projects, has produced the DNA sequences for a large number of organisms, from humans and mice, to zebrafish, yeast, and over eighty bacteria. There has been a great deal of work carried out in applying signal processing and statistical methods to DNA recently(Anastassiou 2001, Yu *et al.* 2001, Anastassiou 2000, Yu and Anh 2004).

In the field of DNA analysis, techniques such as the discrete Fourier transform (Anastassiou 2001) and multifractal analysis (Yu *et al.* 2001) have been explored. This chapter contains new applications of these methods to the areas of sequence analysis (Fitch and Sokhansanj 2000) and phylogenetic trees, those showing the relationships between organisms (Brown *et al.* 2001), respectively. This chapter also contains two complex systems approaches, mutual information and fractals, that can be used to analyse mutations. The application of these approaches are in Chapter Three, in order to analyse a method for modelling mutations in genes.

2.1.1 Novel contributions

The novel contributions contained in this chapter are:

1. Combination of multifractal measure distances with minimal-span tree algorithm to generate phylogenetic trees of bacteria, as compared with the correlation-based, neighbour-joining method used by Yu and Anh (2004).

2. Discovery of a microsatellite region in *Staphylococcus aureus* (Mu50 strain) using a colour spectrogram, generated by a discrete Fourier transform-based method.

3. Developing a way of applying the Higuchi fractal measure to genetic sequences.

2.2 Introduction to DNA analysis

Genetics is concerned with the physical characteristics of organisms that are passed on from one organism to another through the use of deoxyribonucleic acid (DNA), consisting of a sequence of nucleotides on two strands. The nucleotides are the chemical bases adenosine, thymine, cytosine, and guanine that are denoted using the alphabet $\{A, T, C, G\}$. Those on one strand are paired in a complementary fashion with those on the other strand, where adenosine matches with thymine, and guanine with cytosine. Groups of three bases are called codons, and these encode the twenty amino acids that combine to form proteins, the building blocks of life. In a nutshell, the central dogma of molecular biology states that "DNA makes RNA makes protein". This is encapsulated in Figure 2.1. The DNA is transcribed into complementary messenger ribonucleic acid (mRNA). In RNAs, the alphabet is $\{A, T, U, G\}$ where uracil plays the same role that cytosine does in DNA, as it pairs with guanine. Sections of the mRNA that do not code for proteins are removed, and a "poly-A tail"—a sequence composed entirely of adenosine bases—is added to (chemically) stabilise the sequence. The mRNA then acts as a template for protein synthesis. Transfer RNAs (tRNAs) bind to an amino acid on one end, and a complimentary set of three bases on the mRNA template. A 1D sequence of amino acids forms and is then detached from the tRNAs and folds into a 3D structure. This sometimes occurs by itself and sometimes with the aid of other proteins, either immediately or at a later date in the life of the cell. DNA that binds to an mRNA sequence is complimentary to this sequence and is explicitly called cDNA. This principle is used in microarray technologies as described later.

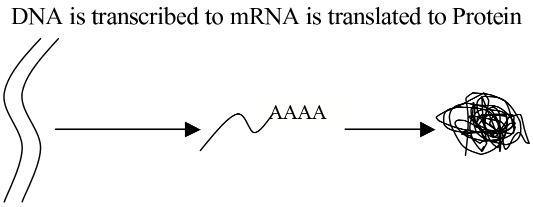


Figure 2.1. The central dogma of biology. The central dogma of molecular biology states that "DNA is transcribed into messenger RNA, which is then translated into protein." This diagram also shows DNA replication, which is carried out with the aid of a number of proteins. At the mRNA stage, introns are spliced out from the sequence, leaving only the protein coding exons. This dogma is of course vastly simplified, for example there is added complexity through splicing, RNA-only genes, RNA-RNA interactions, prions, and other details (Nature 2002, Caporale 2003). But in its essential form this does describe the flow of information in a cell.

Not all regions of DNA code for proteins—some of these non-protein-coding regions have known functions, such as the *XIST* gene (Clerc and Avner 1998), which codes for a ribonucleic acid (or RNA) molecule that deactivates one of the two X chromosomes in female mammals. These RNAs may play an important role in the complexity of organisms such as humans (Mattick 2001). There are also promoter regions around genes that act as targets for gene activation or deactivation (Boyd *et al.* 2003). Other non-coding regions appear to only be “junk” DNA left over from the biological past, with little or no use—or perhaps have a yet undiscovered function. Biologists have suggested that “junk” regions may act as a form of isolation between coding regions and may also act as error-robust locations for sexual recombination. This is described further in Harmer *et al.* (2001), where it is conjectured that these effects could be modelled in game-theoretic terms. It is possible that non-(protein)-coding regions in introns could contain information. As a very simple example of this, it has been shown that increasing the intron length can decrease the probability of (or in other words, the final amount of protein) containing the exon immediately after that intron (Bell *et al.* 1998).

Signal processing is the use of mathematical techniques to analyse any data signal. This data could be an image, a sound, or any other sequence of data, such as a sequence of nucleotides. The sequences of interest could be protein coding regions, repeating elements that may be associated with various diseases—such as Huntington’s disease (Rubinsztein *et al.* 1994)—or regions rich in some set of complementary bases, such as A and T, which can give information on evolutionary history including lateral gene transfer in bacteria (Worning *et al.* 2000).

An area where signal processing techniques have enjoyed wide usage is in microarray processing (Fitch and Sokhansanj 2000). In microarray analysis, effects on gene expression (as ascertained through mRNA levels) can be tested, for example the effect of a drug. Two-colour microarrays are a coloured grid of spots (typically one colour for the control, the other for the cells under test) with spot intensity and colour showing the expression levels for the gene associated with that spot. Affymetrix microarrays only consider one gene and a gene control in a paired-spot arrangement. The control spots control for non-specific hybridisation and background signals. The use of only one flourescent dye removes bias caused by differences in fluorescent dye tagging.

The analysis of the sequences produced has come under intense focus as an area where signal processing could be used to solve a number of important problems such as the nature of non-coding DNA and distinguishing coding DNA from non-coding DNA. Methods such as the discrete Fourier transform (Anastassiou 2001, Anastassiou 2000) and multifractal analysis (Yu *et al.* 2001) have been applied to the problem, complementing more traditional techniques that often use hidden Markov models (Durbin *et al.* 1998, Lukashin and Borodovsky 1998); these are detailed later. A good overview of Fourier transform methods and wavelet transforms, not discussed in this chapter, and a more in-depth discussion of cellular neural networks can be found in Zhang *et* *al.* (2002). Here the focus is on other applications of Fourier methods, and also the application of hidden Markov models and other mathematics to general problems in genetics.

Signal processing is not just a human enterprise. Even individual cells process signals in the form of mRNA, protein, and more general chemical levels (Kholodenko *et al.* 2002, Tyson *et al.* 2003, Thattai and van Oudenaarden 2001, Barabási and Oltvai 2004)—for example sugars in the environment. As with conventional computers, cells can be genetically programmed to process signals (Kobayashi *et al.* 2004, Hasty *et al.* 2002, Ozbudak *et al.* 2002). As in electrical circuits, switching elements can be built in, and positive and negative feedback loops are present, enabling a range of behaviours to be "programmed", such as chemical oscillations of a predetermined frequency. Such engineered "gene circuits" could have important applications in gene therapies that modify or augment the existing protein and cellular interactions in an organism.

2.3 Analysis of regions in DNA using spectrograms

Colour spectrograms are a useful tool in visualizing aspects of signals occurring in time. For example, one can see the noise present in the audio recording of the moon landing and then design an efficient filter to remove it, using only the visual information provided in the spectrogram to determine the noise. Colour spectrograms assign a colour (or brightness) value based on the Fourier transform amplitude or phase at a particular location in frequency and in time—a sliding, short, non-overlapping window of time is taken from a longer time sequence of data, and the discrete Fourier transform is evaluated over the short period of time, over a range of frequencies, at the location in time at which that segment is taken.

Distinguishing coding from non-coding regions is an important problem in genetics. Galván *et al.* (2000) have explored entropy-based methods for separating the coding from the non-coding regions. Can colour spectrograms give a visual guide to these regions? The following are some new results using existing methods that further illustrate that this is possible.

Anastassiou (2001) has explored the use of colour spectrograms in analysing DNA sequences. For a sequence of bases numbered $1, \ldots, n, \ldots, N$, he defines the following sequences:

$$x_r[n] = \tfrac{\sqrt{2}}{3}(2u_T[n] - u_C[n] - u_G[n]),$$

$$x_g[n] = \tfrac{\sqrt{6}}{3}(u_C[n] - u_G[n]), \tag{2.1}$$

$$x_b[n] = \tfrac{1}{3}(3u_A[n] - u_T[n] - u_C[n] - u_G[n]),$$

where $u_X[n] = 1$ if the base at position n is X, or zero otherwise. The sequences x_r, x_g, x_b are used in generating red, green, and blue colour components of pixels (the squares) in the spectrograms. The mapping of a base at position n, from the set $\{a, t, c, g\}$ onto the sequences x_r, x_g, x_b is done to maximise the differences between the sequences at that position n, which results in more vivid colourings of the spectrogram. To colour the spectrogram, compute the discrete Fourier transform (DFT)

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-2\pi jnk/N}, \tag{2.2}$$

where $x(n)$ is the sequence of data $(n = 0, \ldots, N-1)$, $j = \sqrt{-1}$, k is the discrete frequency, and $X(k)$ is the discrete Fourier transform at frequency k. For DNA sequences, one must transform the DNA sequence $s(n)$ into a numerical sequence $x(n)$, or in some cases several numerical sequences $x_i(n)$. One such transformation is that used by Silverman and Linsker (1986). To a sequence of bases, denoted by $\boldsymbol{s} = s(1)s(2)\ldots s(N)$, a vector $\boldsymbol{x}(i)$ is assigned to each base $s(i)$,

$$\boldsymbol{x}(i) = \begin{cases} (1,0,0), & s(i) = A, \\ (-1/3, 0, 2\sqrt{2}/3), & s(i) = C, \\ (-1/3, -\sqrt{6}/3, -\sqrt{2}/3), & s(i) = G, \\ (-1/3, \sqrt{6}/3, -\sqrt{2}/3), & s(i) = T. \end{cases} \tag{2.3}$$

So for example the sequence ATG is represented by the sequence of vectors $(1,0,0)$, $(-1/3, \sqrt(6)/3, -\sqrt(2)/3)$, $(-1/3, -\sqrt(6)/3, -\sqrt(2)/3)$. The next step is to compute the power spectrum

$$P(f) = \sum_{c=1}^{3} \left| \frac{1}{N} \sum_{i=1}^{N} x(i)_c e^{-j2\pi if} \right|^2, \tag{2.4}$$

where $x(i)_c$ is the c-th component of $\boldsymbol{x}(i)$, and $j = \sqrt{-1}$. Here, N is the length of the sequence (number of bases). A simpler method is to use indicator functions

$$x(i) = \begin{cases} 1, s(i) = \alpha, \\ 0, \text{otherwise}, \end{cases} \tag{2.5}$$

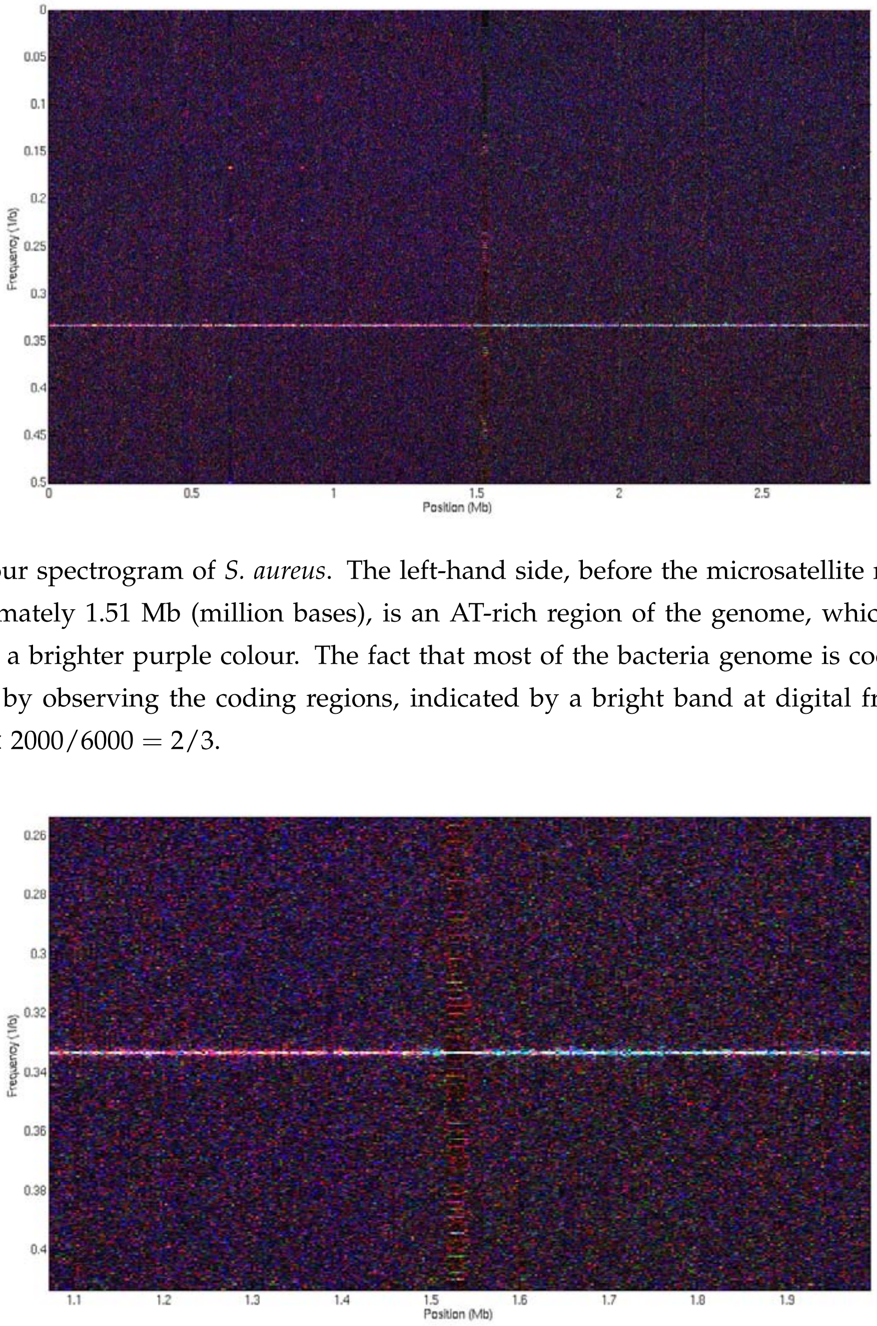
for some $\alpha \in \{A, T, C, G\}$ (Tavaré and Giddings 1989). The power spectra of these two methods are related (Coward 1997),

$$|Y(k)|^2 = \begin{cases} \dfrac{N}{N-1}|X(k)|^2, & k \neq 0, \\ \dfrac{N}{N-1}|X(k)|^2 - \dfrac{c}{N-1}, & k = 0, \end{cases} \tag{2.6}$$

where N is the length of the sequences, c is a constant that varies with N, $X(k)$ is the Fourier transform of the indicator sequence, $Y(k)$ is the average of the Fourier transforms of the sequences of components of the vector sequence as given in Eq. 2.4.

In this work a DFT block size of $N = 6000$ was used since 6000 has a large number of integer dividers, some of which correspond to common repeat lengths in DNA, for example the frequency $k = 2000$ (digital frequency $f_d = 2 \times 2000/6000 = 2/3$) corresponds to the codon length $3 = 6000/2000$. Repeats of length two and six are also common in sections of DNA, these have frequencies of $k = 6000/2 = 3000$ and $k = 6000/6 = 1000$. These repeat lengths thus give rise to centre-cell frequencies, so there is no sidelobe leakage for these repeat lengths.

To illustrate the usefulness of this technique in identifying regions of DNA, Figure 2.2 shows the colour spectrogram of the DNA sequence of the entire *Staphylococcus aureus Mu50* (Kuroda *et al.* 2001) genome. Figure 2.2 clearly shows there are differences in the spectrogram between coding regions of DNA and a microsatellite region at approximately 1.51 Mb in the *S. aureus* Mu50 genome. The lack of fine-grained resolution of the spectrograms is problematic, and prevents easy visualisation of the exact locations of the borders between coding and non-coding regions.



(a) Colour spectrogram of *S. aureus*. The left-hand side, before the microsatellite region at approximately 1.51 Mb (million bases), is an AT-rich region of the genome, which shows up with a brighter purple colour. The fact that most of the bacteria genome is coding can be seen by observing the coding regions, indicated by a bright band at digital frequency $f_d = 2 \times 2000/6000 = 2/3$.

(b) Enlarged version of region in *S. aureus* showing a microsatellite (repeat sequence) region. The repeat length is not an integer divisor of 600, so the power splits into sidelobes, as seen in the bright vertical bands at various frequencies.

Figure 2.2. Colour spectrograms of *S. aureus*..

2.4 Multifractal analysis

One useful way to compare different organisms is by employing a multifractal method. In these, a numerical assignment of bases is used in a set of calculations that leads to a calculation of the Rényi entropy(Rényi 1960). The following is a summary of the fractal method as detailed by Anh *et al.* (2001): to each possible substring $s = s_1 \dots s_k, s_i \in A$ of DNA of length K, there is assigned a unique set, $[x_l, x_r)$, given by

$$x_l(s) = \sum_{i=1}^{K} \frac{x_i}{4^i}, \tag{2.7}$$

where

$$x_i = \begin{cases} 0, & s_i = a, \\ 1, & s_i = c, \\ 2, & s_i = g, \\ 3, & s_i = t, \end{cases} \tag{2.8}$$

and

$$x_r(s) = x_l(s) + \frac{1}{4^K}, \tag{2.9}$$

then

$$F_K(s) = \frac{N(s)}{L - K + 1}, \tag{2.10}$$

where $N(s)$ is the number of occurrences of the substring s in the string of length L of the whole genome. The fractal measure is then

$$\mu_K(dx) = Y_K(x)dx, \tag{2.11}$$

where

$$Y_K(x) = 4^K F_K(s), x \in [x_l(s), x_r(s)). \tag{2.12}$$

The partition sum is defined as

$$Z_\epsilon(q) = \begin{cases} \sum\limits_{\mu(B)\neq 0} [\mu(B)]^q, & q \neq 1, \\ \\ \sum\limits_{\mu(B)\neq 0} \mu(B) \ln \mu(B), & \text{otherwise.} \end{cases} \tag{2.13}$$

Here we run over all non-empty boxes $B = [n\epsilon, (n+1)\epsilon)$ where $\epsilon = 4^{-K}$ and $n = 1, \dots, 4^K - 1$. Since $\mu(B) \in \mathbb{R}$ and addition is commutative in the reals, the ordering of the $\mu(B)$ given by Eq. 2.8 is unimportant in calculating Eq. 2.13. It is therefore unimportant in calculating the Rényi dimension D_q for $q \in \mathbb{R}$, given by

$$D_q = \begin{cases} \lim\limits_{\epsilon \to 0} \dfrac{\ln Z_\epsilon(q)}{(q-1)\ln \epsilon}, & q \neq 1, \\ \\ \lim\limits_{\epsilon \to 0} \dfrac{Z_\epsilon(q)}{\ln \epsilon}, & q = 1. \end{cases} \tag{2.14}$$

Note that although the method used by Yu and Anh (2004) does not show long-range correlations in the DNA sequence, here it is the information content in the sequence that is of interest, and not the correlations. If one considers the case where $q = 1$, then the Rényi dimension D_1 is the same as the Shannon entropy (Shannon 1993). As differences and similarities in G+C content can indicate relationships between organisms (da Silva *et al.* 2002), here the Rényi dimensions are used to determine if this is reflected in a useful way in an uneven distribution of the segments—the ordering is unimportant here, since only the unevenness of the distribution is being compared, not properties relating to the ordering.

The multifractal $D(q)$ plot for *Campylobacter jejuni* (Parkhill *et al.* 2000) is shown in Figure 2.3. As with the Yu and Anh (2004), it was found though trial and error that a segment size of $K = 8$ works best in classifying bacteria. The near linearity of the $D(q)$ plot around $q = 0$ suggests that one can assign to each bacteria a point in $\mathbb{R}^2$ or $\mathbb{R}^3$ given by (D_{-1}, D_1) or (D_{-1}, D_1, D_{-2}). Yu and Anh (2004) found that phylogenetically close bacteria are close in the two spaces. Here the space (D_{-1}, D_1, D_{-2}) is used in conjunction with the minimal-span tree algorithm (Winter 1987) to generate phylogenetic trees in the following subsection.

2.5 Phylogenetic trees

A phylogenetic tree is a tree showing relationships between organisms, including putative ancestral relationships. For each pair $(\boldsymbol{x}, \boldsymbol{y})$ of genomes, one computes the vectors in Euclidean $\mathbb{R}^3$ space

$$\boldsymbol{r}_x = (D_{-1}(\boldsymbol{x}), D_1(\boldsymbol{x}), D_{-2}(\boldsymbol{x}))\,, \tag{2.15}$$

and

$$\boldsymbol{r}_y = (D_{-1}(\boldsymbol{y}), D_1(\boldsymbol{y}), D_{-2}(\boldsymbol{y}))\,. \tag{2.16}$$

Then compute the metric

$$d_{xy}^2 = \|\boldsymbol{r}_y - \boldsymbol{r}_x\|^2, \tag{2.17}$$

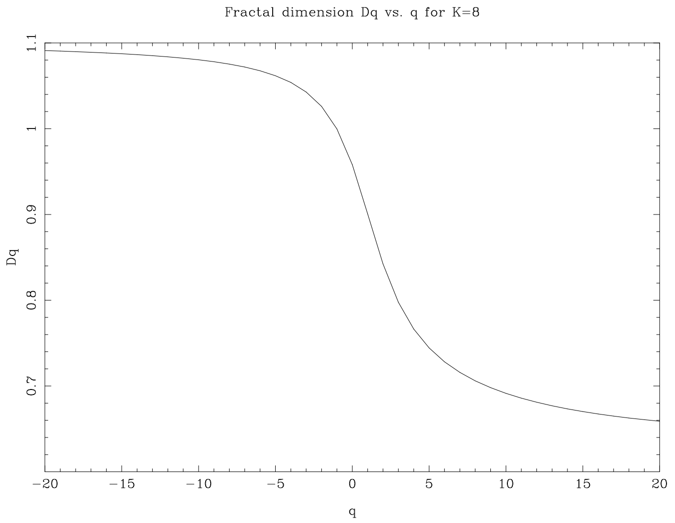


Figure 2.3. Multifractal plot. This is the Rényi (multifractal) dimension plot, with $K = 8$, for the bacteria *C. jejuni*. Note that the value D_1 is the Shannon entropy of the genome for a symbol size of 8. The graph is relatively linear in the region $(-2, 1)$ which suggest these values of D_q can be used as elements of vectors in a Euclidean space.

and this metric can be used in the minimal-span tree algorithm (Winter 1987) to generate binary phylogenetic trees. This approach was taken to generate the phylogenetic tree for members of the proteo-bacteria and hyperthermophile families of bacteria as shown in Figure 2.4(a).

Another method explored herein in relation to both text and DNA is a quantitative chi-squared method that computes a metric with lower scores indicating a closer match. Similar to the inter-word spacing technique for text (Berryman *et al.* 2003a), in analysing DNA one can compute a scaled standard deviation of spacing, in this case for codons. For example, the spacing for the codon *gat* in the sequence *gat agg gcg gat* is two. Note that here a sequence is broken into non-overlapping, adjacent groups of three bases, starting at the beginning, to form codons; while not correct in the sense of the true biology of gene reading, where introns and different reading frames (start positions) occur, these problems are ignored as we are only interested in large scale properties of the sequence. The scaled standard deviations are computed by

$$\hat{\sigma} = \frac{1}{\bar{x}} \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}, \tag{2.18}$$

where the x_i's are the set of spacings for a triplet i.

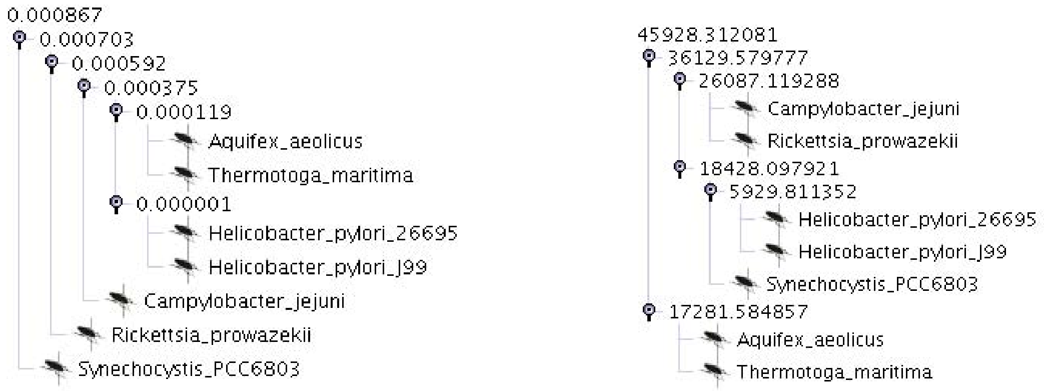
This gives sets of variances of codon spacings for all the genomes, $\{\hat{\sigma}_{11}^2, \ldots, \hat{\sigma}_{I1}^2\}, \ldots,$ $\{\hat{\sigma}_{1J}^2, \ldots, \hat{\sigma}_{IJ}^2\}$, for all possible codons, labelled $i = 1, \ldots, M$ and genomes $j = 1, \ldots, J$. Then the formula for χ^2 as given in Kullback (1968) is used on a pair of genomes $(k, l) \in \{1, \ldots, J\} \times \{1, \ldots, J\}$,

$$\chi_{kl}^2 = \frac{1}{N_k N_l} \sum_{i=1}^{I} \frac{\left(N_l \hat{\sigma}_{ik}^2 - N_k \hat{\sigma}_{il}\right)^2}{\hat{\sigma}_{ik}^2 + \hat{\sigma}_{il}^2}, \tag{2.19}$$

$$N_k = \sum_{i=1}^{I} \hat{\sigma}_{ik}^2, \tag{2.20}$$

$$N_l = \sum_{i=1}^{I} \hat{\sigma}_{il}^2. \tag{2.21}$$

This generates a set of χ^2 values for each pair of genomes. As with the multifractal metric, the chi-squared values can be combined with the minimal-span tree algorithm to produce a phylogenetic tree. For comparison between the trees generated between the metric, see Figure 2.4



(a) Phylogenetic tree obtained using the multifractal metric

(b) Phylogenetic tree obtained using the chi-squared metric

Figure 2.4. Phylogenetic trees of bacteria using the multifractal distance metric. The result of applying the minimal-span tree algorithm to the multifractal distance metric in Eq. 2.17 is shown for several members of the proteo-bacteria family in Figure 2.4(a). Using the chi-squared metric in Eq. 2.19 instead results in the tree shown in Figure 2.4(b). The miniature bug icons represent the organisms currently in existence, the circles represent the branches of the tree (where the software calculates that the species diverged), and the numbers represent the metric scores used to separate the families of bacteria at that point. Clearly the two *H. pylori* (Alm *et al.* 1999) strains group together correctly for both metrics. A comparison with trees obtained by a detailed analysis of proteins (Brown *et al.* 2001), indicates the *Thermatoga maritima* (Nelson *et al.* 1999) and *Aquifex aeolicus* (Deckert *et al.* 1998) as also closely related, and indeed these group together in the phylogenetic trees generated. Of the two trees, the one using the chi-squared metric appears qualitatively more correct when compared with ones generated from the more usual metrics and tree algorithms used in the study of phylogenetic relationships (Brown *et al.* 2001, Snel *et al.* 1999).

2.6 Exploring correlations and mutual information in DNA

2.6.1 Mutual information functions

Another method for showing the existence of long-range correlations in DNA is to use the mutual information function, as given in Eq. 2.22 below. This approach has been shown to distinguish between coding and non-coding regions (Li 1992). In Chapter Three, an application is given for the mutual information function in Eq. 2.22:

$$M(d) = \sum_{\alpha \in A} \sum_{\beta \in A} P_{\alpha\beta}(d) \log_2 \frac{P_{\alpha\beta}(d)}{P_\alpha P_\beta}, \tag{2.22}$$

for symbols $\alpha, \beta \in A$ (in the case of DNA, $A = \{a, t, c, g\}$). $P_{\alpha\beta}(d)$ is the probability that symbols α and β are found a distance d apart, and P_α and P_β are the probabilities of finding symbols α and β at any location. The mutual information function is related to the correlation function (Li 1990):

$$\Gamma(d) = \sum_{\alpha \in A} \sum_{\beta \in A} a_\alpha a_\beta P_{\alpha\beta}(d) - \left(\sum_{\alpha \in A} a_\alpha P_\alpha \right)^2, \tag{2.23}$$

where a_α and a_β are numerical representations of symbols α and β. As discussed by Li (1990), the fact that we are working with a finite sequence means that this $M(d)$ overestimates the true $M_T(d)$ by

$$M(d) - M_T(d) \approx \frac{K\,(K-2)}{2N}, \tag{2.24}$$

where K is the number of symbols (for DNA this is always 4) and N is the sequence length. The shortest sequence used was the sequence of the *Homo sapiens* immunoglobulin superfamily, member 8 gene (GenBank accession BC004108), which was $N = 1750$ base pairs in length. Thus for this gene the difference between the estimated and real mutual information is $\approx \frac{4\times 2}{2\times 1750} = 0.002$, which is a factor of ten less than the mutual information estimate for this gene. Furthermore, since the results below are comparing the mutual information of the sequence with that of the random sequence, the inaccuracy is effectively eliminated, since the error should be roughly the same for these sequences of the same length N and of course the same $K = 4$.

The mutual information is, at least for large d, proportional to the correlation squared, $\Gamma^2(d)$ (Li 1990). Even for small d, the mutual information function still provides an estimate of the correlations. The range of d used (up to 1024) means that a reasonable estimate of the correlations at these larger distances is provided herein. In biological terms, correlations within regions of genes, and between promoter regions and DNA are being captured. This length is not sufficiently large to explore longer range correlations such as those between genes—typically tens of thousands of bases—or those that might exist between activator or silencer regions and promoters, again on the order of tens of thousands of bases (Levine and Tjian 2003). In whole chromosome analysis one would expect to find repeating elements and other correlations in "junk" DNA in addition to correlations within genes.

2.6.2 Higuchi fractal measure

A method for determining correlations in sequences is the Higuchi fractal method (Higuchi 1988). Using this method one can compute the Higuchi fractal measure

$$L\left(k\right) = \sum_{m=0}^{k-1} \frac{N-1}{\lfloor \frac{N-m}{k} \rfloor k^2} \sum_{i=1}^{\lfloor \frac{N-m}{k} \rfloor} \left| x\left(m+ik\right) - x\left(m+\left(i-1\right)k\right) \right|, \tag{2.25}$$

for $k = 1, \ldots, 1024$ over non-overlapping subsequences of length 4000. The sequence $x\left(i\right)$ is generated by mapping the sequence of bases, $s\left(i\right)$:

$$x\left(i\right) = \begin{cases} 1.0, & s(i) = a, \\ 0.5, & s(i) = t, \\ -0.5, & s(i) = c, \\ -1.0, & s(i) = g. \end{cases} \tag{2.26}$$

Performing linear regression on $\log L\left(k\right)$ versus $\log k$ then gives a slope of $-D$, where D is the estimate of the true fractal measure. For a high degree of correlation, one would expect a value of D closer to one.

One can also apply the Higuchi method to the density of bases in blocks, as carried out by Lu *et al.* (1998), however this does not provide a measure of correlations in the sequence as the authors claim, but rather correlations in the density function. In the fashion in which it is used in Chapter Three, it can, like the mutual information function, detect correlations in the actual sequence. As described above it detects correlations up to 1024 base pairs apart.

2.7 Conclusions

This chapter presents a number of interesting methods for analysing DNA. The colour spectrogram method shows promise in highlighting coding versus non-coding regions, microsatellite repeat regions, and AT (or GC) rich regions. Multifractal analysis shows a rich amount of information when applied to whole genomes, and this was successfully used to classify bacteria. Results on the mutual information and Higuchi fractal method are left for Chapter Three.

To summarise the novel contributions of this work, they were the combination of multifractal measure distances with the minimal-span tree algorithm to generate phylogenetic trees of bacteria, the discovery of a microsatellite repeat sequence (using a colour spectrogram method) in a strain of *S. aureus*, and extending the Higuchi fractal measure to genetic sequences.

Chapter 3

Mutations

THIS chapter examines two methods from Chapter Two for determining whether long-range correlations exist in DNA: a fractal measure and a mutual information technique. The performance of these methods and implications of the results are examined in detail. They are used to compare DNA sequences from a variety of sources. Using software for performing *in silico* (simulated) mutations, evolutionary events leading to long-range correlations are considered and analysed using the techniques presented in Chapter Two. Comparisons are made between these virtual sequences, randomly generated sequences, and real sequences. Correlations in chromosomes from different species are also explored.

3.1 Introduction

DNA is a structure containing a long sequence of complimentary pairing bases, denoted by the symbol set $\{a, t, c, g\}$ (Watson and Crick 1953, Franklin and Gosling 1953). The genetic material in DNA undergoes a variety of different mutational events (Joset and Michel 1993, Dover 2000). These mutational events can be considered as string rewriting rules (Durbin *et al.* 1998) that lead to correlations in DNA. Repeated use of short sequences as promoters (Levine and Tjian 2003), or as intron markers (Bon *et al.* 2003) can give rise to very long-range correlations.

A number of different techniques have been studied for examining long-range correlations in DNA. These include Lévy walks (Peng *et al.* 1992), Fourier transforms (Li and Kaneko 1992, Anastassiou 2000, Anastassiou 2001), and wavelets (Arneodo *et al.* 1995). A number of researchers have attempted to explore this by considering power law relationships in power spectra of DNA sequences. This purports to show long-range correlations and also to show differences between regions of DNA. In this chapter long-range correlations are explored using a mutual information technique (Li 1990), and the Higuchi fractal method is briefly explored (Higuchi 1988).

DNA sequences contain a number of coding regions. These are regions that code for protein and are marked with stop and start codons, although the presence of these does not necessarily indicate a coding region. Coding regions may contain introns, which are regions that get spliced (cut) out before translation from the RNA template before the protein is made according to the code on the RNA template (which in turn comes from the DNA). Non-coding regions may just be junk, or may code for regulatory RNAs (Mattick 2001), such as the *XIST* gene which switches off the extra X chromosome in women (Avner and Heard 2001).

This chapter shows that long-range correlations exist for real sequences of DNA and virtual sequences of DNA, but not random sequences of DNA. The virtual sequences of DNA are produced by software that simulates a variety of mutational events. The random DNA has a random sequence generated in software, so it should contain almost no correlations. Also explored is whether or not the power spectra show any differences between coding and non-coding DNA, and between different species of bacteria.

3.1.1 Novel contributions

The novel contributions of this work are:

1. Development of software for performing *in silico* mutations covering a wide variety of scales, but ignoring structure (Karchin *et al.* 2005) and/or fitness (Dasgupta *et al.* 2003).

2. Showing how repeated mutations of different types generate (relatively) long-range correlations in DNA.

3. Showing that the Higuchi fractal measure provides useful information on correlations in DNA.

4. Showing a relationship between the Higuchi fractal measure and a mutual information measure on gene sequences.

This work has been cited by Dehnert *et al.* (2005), who found synchronisation between related bacteria in correlation structures, and by Sadovsky (2006), who looked at the information capacity of genomes.

3.2 Sequences examined

For exploring correlations at very large distances, the following chromosomes were used: *Homo sapiens* chromosome 20 (Deloukas *et al.* 2001), *Mus musculus* chromosome 2 (Mouse 2002, Gregory *et al.* 2002) and *Escherechia coli* (Hayashi *et al.* 2001) (*E. coli* only has one chromosome).

3.2.1 Real sequences

In order to compare correlations in real DNA with those in short random and short virtual DNA sequences, a selection of twenty short, real gene sequences from various organisms was chosen. Their accession numbers, and descriptions are shown in Table 3.1.

Table 3.1. Details of the real mRNA sequences used. The GenBank (Benson et al. 2003) accession numbers and descriptions of the twenty short, real mRNA sequences used. For a discussion of messenger RNA (mRNA) see Chapter Two

NOTE: This table is included on page 26 of the print copy of the thesis held in the University of Adelaide Library.

3.2.2 Random sequences

To compare the mutual information in real and virtual sequences, twenty random sequences of 10 000 bases in length were generated, where all four bases have equal probability of appearing in each position. This equiprobability is not exactly true in general, but it does not matter for the purposes of generating virtual sequences, what matters is a reasonable model of the types of mutations.

3.2.3 Virtual sequences

The twenty virtual non-coding regions are generated by the latest version of my software for exploring mutations in DNA (Berryman *et al.* 2003b). It implements the following *in silico* operations:

- Base substitutions, where one base pair has been replaced with a different base through some mechanism (such as UV irradiation with an absent or partly unsuccessful repair process).
- Additions, where a base pair has been added to the sequence.
- Deletions, where a base pair has been removed from the sequence.
- Flips, where part of a sequence has been replaced by its reverse complement.
- Fills, where a sequence of repetitive elements (of length 1 to 4) has been inserted up to 50 times. The exact number of repetitions is chosen at random from a uniform distribution, as is the length.
- Copies, where part of a sequence (up to 100 bases in length) has been copied. As with the fill operations, the length is chosen from a uniform random distribution.

The flip, fill, and copy operations are illustrated in Figure 3.1. These operations are meant to simulate small scale general mutations, and larger scale ones of the type that occur in non-coding DNA. A more detailed discussion of the biology can be found in Chapter Six (Section 6.1), but for the purposes of this model, this can safely be ignored, as the base mutation frequencies and specifics of large scale processes do not change the overall trends modelled herein. In each run of the simulator one of the random DNA sequences was taken and up to 30 (the exact number chosen from a uniform random distribution) of each of the above mechanisms were used to generate long-range correlations in the DNA sequences. With some experimentation it was found that, as one would expect, the fill and copy mechanisms are the primary drivers in creating long-range correlations.

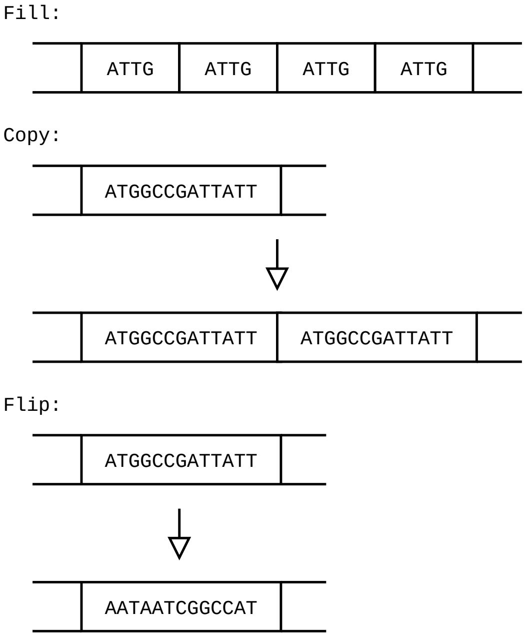


Figure 3.1. Mutation operations. This figure shows the three operations: fill, where a sequence of repetitive elements of length 4 (in this case) is added; copy, where a part of the DNA sequence is copied; and flip, where part of the DNA sequence is replaced by its reverse complement.

3.3 Results

3.3.1 Short DNA sequences

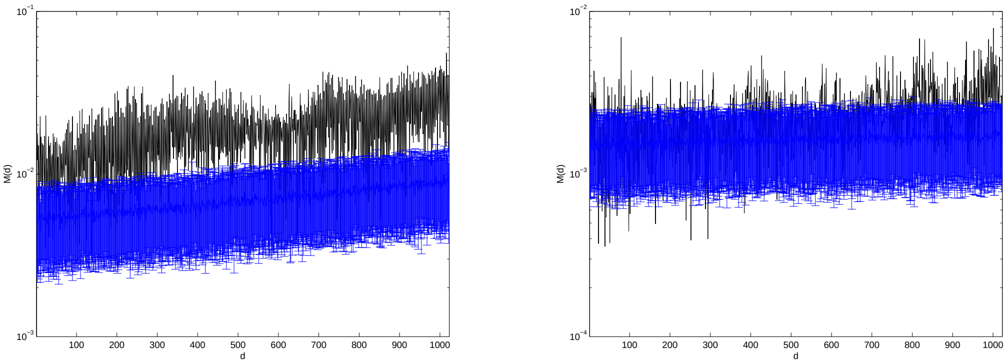
To analyze the short DNA sequences (real, virtual, and random) using the mutual information function (Eq. 2.22), one can compare the mutual information plot with the average ± standard deviation plot of the mutual information function. This is done for 100 randomised sequences with the same base distribution, but in random order, thus eliminating correlations. Examples of this are shown in Figure 3.2.

The maximum distance at which significant correlations were present was determined, up to the maximum distance studied of 1024. The results of this for the 20 real, virtual, and random sequences are shown in Table 3.2. No long-range correlations are present in the benchmark random sequences as one would expect, however correlations up to distance $d > 1024$ are present in the virtual sequences, and even longer range correlations of distance $d > 1024$ can be found in real sequences. Because the mutation process used to generate the virtual sequences was random, there was a significant variation in the length of correlations present. This corresponded well to the number of repeated elements and copy mutations, in particular with the copy mutations. Future work will attempt to quantify the mutual information values with a directed model of evolution where real sequences are taken and mutation operators applied in a realistic fashion. For example, point mutations are much more likely to be seen in the "wobble" positions of codons than elsewhere, and this in turn is much more likely than insertions and deletions.

The results of using the Higuchi fractal method are shown in Table 3.3. Note that these estimates are relatively independent of the choice of mapping of bases onto numbers—several different mappings were tried with variations on the order of 0.001—and the numbers are in fact overestimates of the true fractal dimension. The fractal dimensions appear unrelated to the mutual information distances, thus illustrating the fact that the mutual information function is a better characterization of the distances at which correlations are present.

Table 3.2. Approximate distance in base pairs at which there is no significant mutual information. This table shows the approximate (± 50) distances at which the mutual information function drops down to the level of the uncorrelated sequences of the same base distribution. The numbering of the real sequences matches the ordering they are given in Table 3.1. The numbering of the virtual sequences corresponds to the random sequence which was mutated to produce that virtual sequence, but bears no relationship to the numbering of the real sequences.

Sequence number	Random	Virtual	Real
1	0	0	> 1024
2	0	0	> 1024
3	0	> 1024	700
4	0	100	800
5	0	50	0
6	0	0	> 1024
7	0	850	> 1024
8	0	0	> 1024
9	0	0	> 1024
10	0	800	> 1024
11	0	0	> 1024
12	0	> 1024	> 1024
13	0	100	950
14	0	> 1024	600
15	0	> 1024	> 1024
16	0	0	> 1024
17	0	0	> 1024
18	0	0	> 1024
19	0	> 1024	> 1024
20	0	0	> 1024



(a) This figure shows the plot of the mutual information function $M(d)$ in Eq. 2.22 against base distance d for the sequence of the MAP kinase-activated protein kinase 2 gene from *Mus musculus*, shown in a darker line style, compared with the set of 100 randomised sequences of the same base distribution, the lighter band. The graph of mutual information in the MAP kinase gene mostly sits about the "noise floor" of the randomised sequences, in which the correlations have been destroyed.

(b) This figure shows the plot of the mutual information function $M(d)$ in Eq. 2.22 against base distance d for the virtual DNA sequence number 14, shown in a darker line style, compared with the set of 100 randomised sequences of the same base distribution, shown as a lighter band. The graphs mostly overlap, indicating few significant correlations in the virtual sequence when compared with the randomised sequences containing little to no correlations.

Figure 3.2. Plots of mutual information for real and virtual DNA sequences. This figure show the plots of the mutual information function $M(d)$ in against base distance d for (a) a real sequence and (b) a virtual sequence. At larger distances, there are fewer symbols at that distance that are available for computing the mutual information, so the over-estimates increase in value, producing a slight slope to the graphs

Table 3.3. Estimate of fractal dimension of DNA sequences. This table shows the estimates of the fractal dimension as ascertained using the Higuchi method described by Eq. 2.25. The numbering of the real sequences matches the ordering they are given in Table 3.1. The numbering of the virtual sequences corresponds to the random sequence which was mutated, but bears no relationship to the numbering of the real sequences.

Sequence number	Random	Virtual	Real
1	1.104	1.103	1.098
2	1.103	1.094	1.095
3	1.104	1.094	1.118
4	1.104	1.086	1.110
5	1.103	1.086	1.092
6	1.102	1.094	1.103
7	1.105	1.100	1.105
8	1.103	1.102	1.087
9	1.102	1.093	1.080
10	1.103	1.099	1.099
11	1.103	1.089	1.087
12	1.104	1.103	1.098
13	1.103	1.099	1.098
14	1.104	1.091	1.055
15	1.104	1.100	1.101
16	1.103	1.103	1.090
17	1.102	1.102	1.097
18	1.102	1.091	1.094
19	1.102	1.099	1.099
20	1.103	1.099	1.091

Table 3.4. Average Higuchi fractal dimension over whole chromosomes. This table shows the average Higuchi fractal dimension D over blocks of length 4000 in the chromosomes listed, along with the variance, and the distance d at which correlations exist as determined by mutual information function in Eq. 2.22

Sequence	$\text{mean}\,(D)$	$\text{var}\,(D)$	d
Eschercia coli K12, complete genome	1.10039	2.07×10^{-5}	> 1024
Mus musculus chromosome 2	1.09691	7.59×10^{-5}	> 1024
Homo sapiens chromosome 20	1.089	0.00991	> 1024

3.3.2 Whole chromosome sequences

The results of analyzing chromosomes from *E. coli*, *M. musculus*, and *H. sapiens* using both the Higuchi fractal measure, D, and the mutual information function, $M(d)$, indicate the presence of correlations up to the maximum length explored (1024). This is shown in Table 3.4. There is less variation in these measures for *E. coli*, which has a greater proportion of gene-coding DNA than other sequences. These gene-coding regions allow less room for repeating elements due to evolutionary and size constraints, and thus have a lower correlation distance.

3.4 Conclusions

This Chapter showed that long-range correlations are present in short sequences of real DNA, "virtual" DNA, and throughout whole chromosomes. The simulation of genetic mutation events in "junk" DNA with fill, copy, and mutate operations also produces long-range-correlations approaching 1024 bases in length. The negative test, with computer generated random sequences, succeeds in that no significant long-range correlations were found. These results confirm that mutational events in non-conserved regions of DNA can give rise to long-range correlations.

To summarise the novel contributions of this work: they are the development of software for performing *in silico* mutations on a variety of scales, how these mutations generate correlations, and relationships between the Higuchi fractal and mutual information measures in studying these correlations.

Chapter 4

Viruses and memes

IN this chapter a variety of network models describing transmission across a network are explored. In particular there is a focus on transmission across composite networks, or “networks of networks”, in which two or more separate networks are interconnected.

In a disease context, the introduction of two interrelated viruses to hosts in a network is simulated, in order to model the infection of hosts in a classroom situation, with high rates of infection within a classroom due to close contact of longer duration, and lower rates of infection between classrooms.

The hosts can be either **S**usceptible to infection, **I**nfected, or **R**ecovering from each virus (an SIR model). During the infection stage and recovery stage there is some level of cross-immunity to related viruses. The effects of immunizing sections of the community are explored

In a share market context, memes, or “viral ideas”, are introduced into a virtual agent-based model of a share exchange. By varying the parameters of the individual traders and the network structure, emergent behaviour can be demonstrated, including boom and bust cycles.

4.1 Introduction

4.1.1 SIRS model

An SIRS "household" model describes the spread of an infectious agent through a network with two types of connections between infectable agents (Ball 1996):

1. Local connections within a household that have a high probability of infection transmission.
2. Connections between households that have a low probability of infection transmission.

The three states of an SIR model are:

1. Susceptible to infection.
2. Infected.
3. Recovered or removed (by death or quarantine).

The agents modelled in this chapter are people, but they could be any organism susceptible to any type of infection. In an SIRS model, agents then move back into a susceptible state due to the introduction of new strains of the virus from an external source or to the existing virus evolving within the network. In the model presented, hosts are not permanently removed through death as only the common cold is considered as passing through an otherwise healthy population. Further, since a rapidly mutating virus is considered, hosts are not considered as removed through having permanent immunity to a virus. Quarantine is also not considered, however this is not as good an assumption as the previous two, since in the data compared with, there are a number of students absent when sick due to their infection.

SIS (where the R state is ignored) and SIR household models have been used to analyze infections within sexual networks (Liljeros 2004, Jones and Handcock 2003) and computer networks (Leveille 2002). Typically, mean field analysis is used to determine the behaviour of the systems under varying parameters (Ball 1999, Ball 2001, Ghoshal and Sander 2004). More recent analytical work by Rand (1999) and Keeling and Rand

(2001) takes a stochastic differential equation approach. Specifically, the authors develop correlation equations which describe the time evolution of low-order correlations. The particular form of the correlation equations are called pair approximations because they are stochastic differential equations for the second-order moments (pair numbers). These better capture the spatial nature of epidemics in which local correlations matter. A related perspective is the work of Koplik *et al.* (1988), which covers transport and dispersion in random networks. One could consider the initial infection of an agent as an "injection" into a random network and consider the spread of the virus in terms of its diffusion and transport through the network.

4.1.2 Memes

The term "meme" comes from Dawkins (1989), and refers to "a unit of cultural transmission, or a unit of imitation." A meme can perhaps best be defined as "a viral idea" (Gaiman 2005), and this is the one used here. In this chapter a meme is considered to be an idea about the value of a company (as reflected in its share price) and these memes are spread through a network. The spread of memes and their effect on the share market has been explored by Frank (1999), who based his memes' values (and hence efficiency of spread via imitation) on the share price return. He found that if the meme value is based on the return, then only long term fluctuations can be observed, with no explanation of shorter term fluctuations (such as those seen in "day trading"). Here these fluctuations are studied, and hence memes' values are based on their immediate effect on generating profits (or losses).

4.1.3 Novel contributions

The novel contributions of this work are:

1. Development of a model for examining the transmission of viruses in a primary school, with a variety of different social interactions: for example separate science classes and lunchtime versus classroom interaction of teachers and children among other things.

2. Development of a model of social interactions between sharemarket traders, with a spread of viral ideas (memes) about share prices, and exploration of the network structure on the share market (price) fluctuations.

4.2 Model

4.2.1 Viruses

An entire school of children with grades $\{g\}$ and classes for each grade $\{c_g\}$ was modelled. It had several types of (social) network connections:

1. Connections within a classroom, such as in Figure 4.1 where the squares denote students, along which there is a high probability p_1 of infection due to large amounts of time together in close contact.

2. Connections between friends, both within and between classrooms, but not between students in different grades.

3. Connections between a science teacher and all students of each grade that has separate science classes.

4. Connections between all teachers of the school.

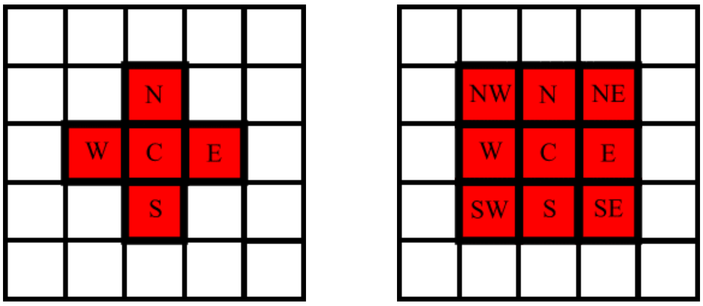
Connection types 2-4 are modelled with the same probability p_2 of infection due to the similar amount of time (a science class or lunch time) spent together.

The type of connection networks used in the modelling herein are either fully connected networks, in that every person within a classroom is connected to every other person, or connected in terms of a Moore or von Neumann neighbourhood as typically used in cellular automata. These are shown in Figure 4.1. When these two types of neighbourhoods are used, the students are arranged in square classes of size $x = \lfloor\sqrt{N_{c_g}}\rfloor$ where N_{c_g} is the size of class c_g. For $N \neq x^2$ an extra row of size $N - x^2 < x$ can be created to make the class square in size. The probabilities used are shown in Table 4.1 and have largely been taken from the literature (Gwaltney 2000, Gwaltney *et al.* 1967). Probabilities p_1 and p_2 have been estimated from the data by solving equations 4.1 and 4.2 simultaneously and repeatedly over the set of times t (thus we gain an understanding of the time evolution of the probabilities). These equations are

$$p_1 \frac{I(t)}{N(t)} + p_2 \frac{I_o(t)}{N_o(t)} - p_1 \frac{I(t)}{N(t)} p_2 \frac{I_o(t)}{N_o(t)} = \frac{I(t+1) - I(t) + R(t+1)}{N(t) - R(t) - I(t)} \tag{4.1}$$

and

$$\begin{aligned} p_1 \frac{I(t+1)}{N(t+1)} + p_2 \frac{I_o(t+1)}{N_o(t+1)} - p_1 \frac{I(t+1)}{N(t+1)} p_2 \frac{I_o(t+1)}{N_o(t+1)} \\ = \frac{I(t+2) - I(t+1) + R(t+2)}{N(t+1) - R(t+1) - I(t+1)}, \end{aligned} \tag{4.2}$$



(a) Von Neumann neighbourhood.

(b) Moore neighbourhood.

Figure 4.1. Neighbourhoods typically used in cellular automata. Cellular automata have the concept of a neighbourhood, which defines the connections between a cell to its neighbouring cells. In the Von Neumann neighbourhood (a), every square at position (x', y') is updated based on the states of the cells in the neighbourhood $\{(x, y) : |x - x'| + |y - y'| \leq 1\}$. In the Moore neighbourhood (b), every square at position (x', y') is updated based on the states of the cells in the neighbourhood $\{(x, y) : \max(|x - x'|, |y - y'|) \leq 1\}$.

Table 4.1. Table of probabilities used for the simulation of school data. These are the probabilities of going between infection states. All variables have a possible maximum range of $0 \leq p \leq 1$.

Variable	Description	Default value/range explored
p_1	Probability of infection being transmitted from a neighbour in the class	x
p_2	Probability of infection being transmitted for other connections	y
p_{IR}	Probability of going from an infectious state to recovered	0.2
p_{RS}	Probability of going from the recovered state to susceptible	0.1
p_f	Probability of a person being friends with another person in the same grade	0.2
η	Individual immunity to a virus	$[0, 1)$ in steps of 0.05

where I is the number of children infected in the class, I_o is the number of children in the other classes in the grade, N is the total number of people in the class, and R is the number of children in the recovery state.

There is a high probability of infection within a class, in line with studies of transmission between children in close contact in a hospital setting (Goldmann 2001).

The total probability of a node being infected by virus i in a single time step is given by

$$p_{SiIi} = \min\left(1, l_{Ii} p_{1i} + f_{Ii} p_{2i}\right)(1 - \eta) \prod_{i \neq j} \left(1 - \left(\delta_{s_j I} + \delta_{s_j R}\right) \alpha_{ij}\right), \tag{4.3}$$

where α_{ij} is the surface protein similarity between viruses i and j, l_{Ii} is the number of local neighbours (within the class) in the infected state for virus i that the node is connected to, f_{Ii} is the number of friends (in other classes) in the infected state for virus i that the node is connected to, and $\delta_{s\mathrm{X}}$ is the Kronecker delta function,

$$\delta_{s\mathrm{X}} = \begin{cases} 1, & s = \mathrm{X} \\ 0, & \text{otherwise.} \end{cases} \tag{4.4}$$

The surface proteins of the virus act as recognition targets for the immune system, and thus having a high α means that acquiring one of the viruses means the immune system has a high probability of recognising the other virus. The matrix $\left[\alpha_{ij}\right]$ is a symmetric matrix with $\alpha_{ji} = \alpha_{ij}\ \forall i \forall j$ and $\alpha_{ii} = 1\ \forall i$.

Levels of immunity η vary with the age of the person. Students within each grade g are treated as having a fixed general level of immunity η_g. The model was compared with data from an actual school with grades K4, K5 and 1-7. Here, K4 and K5 are kindergarten classes containing only four and five year olds respectively, the other grades contain students of a mix of ages (roughly a year apart). Values of immunity used are $\eta_{\mathrm{K4}} = 0.0$, $\eta_{\mathrm{K5}} = 0.05$, $\eta_{\mathrm{teachers}} = 0.6$ and for grades 1 to 7 the immunity is given by

$$\eta_x = 0.05 + \tfrac{x}{16}, \quad x \in 1, \ldots, 8, \tag{4.5}$$

where x is the grade, which seems to be a fit for the school data provided.

4.2.2 Memes

The model market that the memes influence consists of a number of agents (traders), each with the following attributes:

- A bank balance in cents, that is discretised in units of cents.
- A portfolio of shares, detailing how much share of each company the agent holds.
- A set of memes, with a maximum of one per listed company per agent.

Each meme contains the following information:

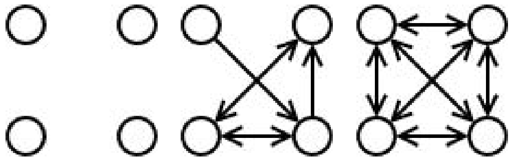
- Stock price in cents, discretised in units of cents. There is a minimum price of 1c and no maximum price. The initial share price is set to a random number (uniform distribution) in the interval, $\left[\max\left(1, S_{\mathrm{sharename}}\left(0\right) - 100\right), S_{\mathrm{sharename}}\left(0\right) + 100\right]$ where $S_x\left(t\right)$ is the share price of share x at time t.

- Volume of shares to trade in. This is initially assigned a number in the range 0-1000 inclusive, at random with a uniform distribution.
- A counter c to keep track of how many successes the meme has generated for the agent in the share exchange. This is incremented each time an agent makes a profit and decremented each time it makes a loss. It has a minimum value of zero.

In order to transmit memes, three things need to be defined: a network structure, a probability of transmission of meme along links in the network, and the specific mechanism of copying. Initially, a simple network structure was considered where the network nodes (traders) were initially connected via directed links with probability p of a directed link from node A to node B. The memes spread only in the direction of the link. The set of p values used for networks were $\{0, 0.5, 1\}$ for networks consisting of a single subgroup of 300 agents. These are illustrated in Figure 4.2.2. Also considered is a scenario where the set of 300 agents was divided into two halves and one half was connected with probability $p_{11} = 0.5$, the other group with probability $p_{22} = 1.0$ (that is, fully connected) and connections from members of the first to the second group and second to the first group with probabilities $p_{12} = p_{21} = 0.2$. The copying is done by replacing an existing meme for share x with a new meme copied along the link with probability

$$p(c) = \frac{1}{1 + \exp(6 - c)}, \tag{4.6}$$

where c is the counter of the success of the meme as defined above. The function is a simple sigmoid function, and is thus constrained to be in the range $[0, 1]$ as is required for a probability, and is a low probability for low values of c and high for high values of c. The $(6 - c)$ term is there to shift the function such that it is almost zero probability for a zero counter, $c = 0$.



(a) $p = 0$ (b) $p = 0.5$ (c) $p = 1$

Figure 4.2. Networks of share exchange traders. This shows a very simple network of four share exchange traders, with varying probabilities p of the formation of directed links between pairs of traders in the initial network construction.

4.3 Methods

4.3.1 SIRS model

To ensure the model captures important aspects of real infections in a school situation it was compared with actual school data from the Colegio Nueva Granada in Bogota, Colombia, collated by the Universidad de los Andes. The data consisted of reports of infections in students from grades (that is years, or forms) 1-7 and K4 and K5. Given the duration and severity of infections, it is most likely the infections were caused by rhinoviruses (Gwaltney 2000, Gwaltney *et al.* 1967) and not bacteria or the influenza virus (Goldmann 2001). Further, the common cold is much more often (30-50% of cases) caused by a rhinovirus than an influenza virus—around 5-15% of cases (Heikkinen and Järvinen 2003). The classes were approximately 20 students and there were on average 6 classes per grade. The model directly produces information on the number of infections per person per ten week period and the numbers of people in each state in each time step.

4.3.2 Memes

My computer simulated the share market for 100 time steps (or "ticks"), using 300 agents. In the first simulation, each agent was given a random bank balance between 1c and $100. Since inflation was observed and this obscured the other trends (as discussed further below), the bank balance was restricted to be in the range 1c to $10 in order to contain inflationary pressure. To simulate different dynamics, various connection probabilities (including having two subgroups) were used to observe the effect of network structure and the spread of memes in different networks.

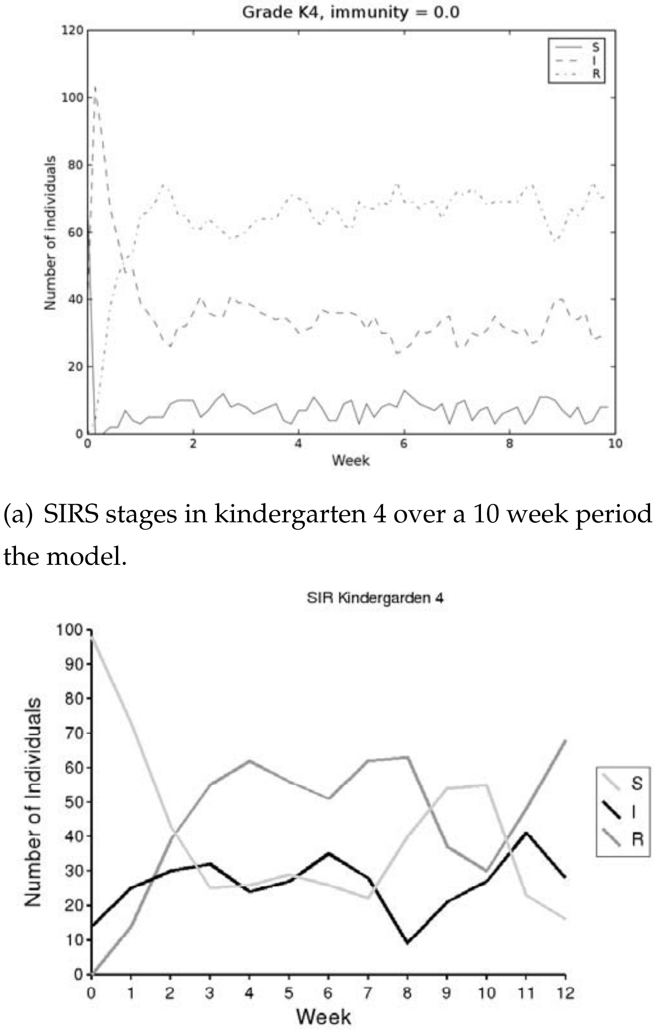
4.4 Results

4.4.1 SIRS model

Figures 4.3 to 4.6 show the number of people in each state over a 10 week period, from the data and the model with parameters chosen so that the plots are as close (by visual inspection) as possible. A fully connected network is used, and there is only one virus being spread through the network. As you can see, the results are similar but different in many ways. This could be due to a number of different causes such as inability to easily quantitatively match the model parameters to the real data set, the inability to assess the true network structure, or simply the stochasticity of the model.

4.4.2 Memes

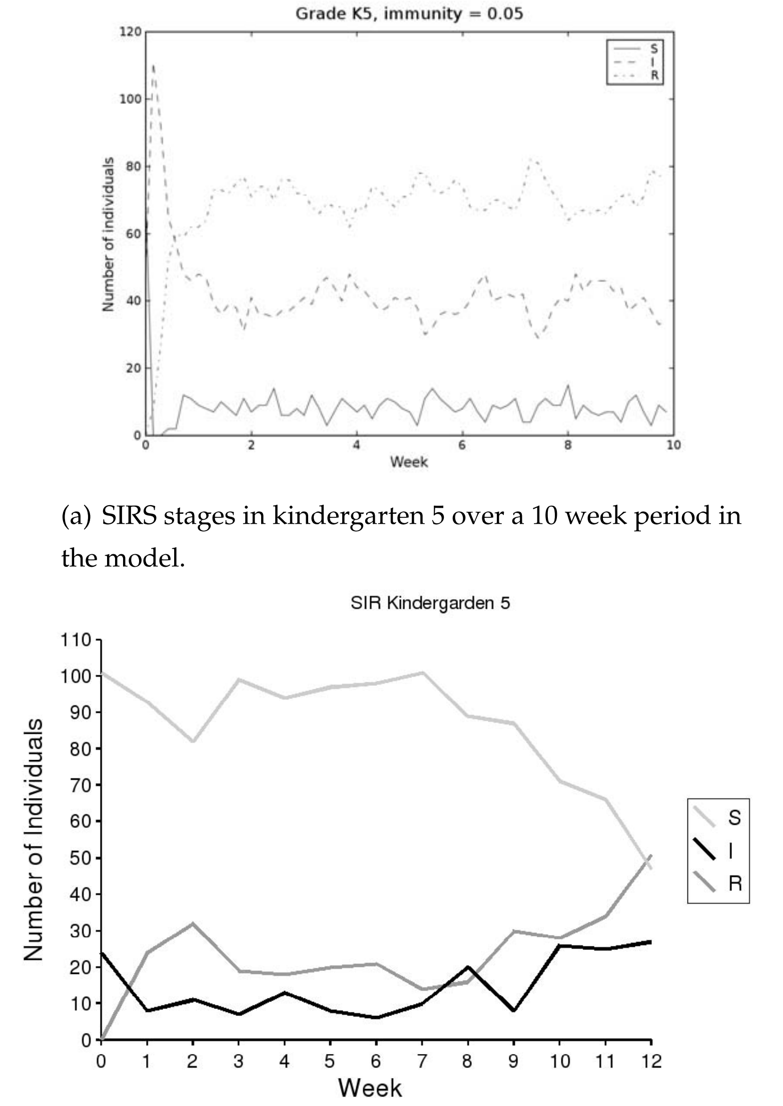
The results with a connection probability $p = 0.5$ and bank balances set at random (uniform) in the range 1c–$100 are shown in Figure 4.7. One can clearly see inflation occurring, due to a high level of free cash injected at time $t = 0$. Note that although there are three shares used in all the simulations, only results for the MSFT share (not the one on the New York Stock Exchange) are shown. The range was reduced to 1c–$10 and the connection probability was varied from $p = 0$ (Figure 4.8) to $p = 0.5$ (Figure 4.9) and finally $p = 1$ (Figure 4.10). Comparing the one at $p = 0.5$ to the inflationary scenario demonstrates how reducing the amount of free cash in the economy significantly reduces inflationary pressure, and allows other stochastic meme effects to be more visible. When the connection probability is reduced to 0, there is no spread of memes and therefore no chance of any boom effects. When this is increased to 0.5 a boom effect is noticable, and finally with a connection probability $p = 1$ we are seeing some interesting dynamics with higher boom effects but also some significant falls in the share price. The change from one main group to two subgroups of traders further as described in Subsection 4.2.2 accentuates the different dynamics possible in the spread of memes, and competition between memes gives rise to the sharp boom and bust cycles shown in Figure 4.11. A diagram showing the spread of memes can be seen in Figure 4.12. This shows that the different subgroups have different dynamics. Investigation of the time-changes of the distributions of memes (not shown) reveals that one meme usually builds in popularity in one subgroup before spreading to the other subgroup, by which time the meme has already evolved in the first subgroup.



(a) SIRS stages in kindergarten 4 over a 10 week period in the model.

(b) Actual data of SIRS stages from kindergarten 4 over a 10 week period.

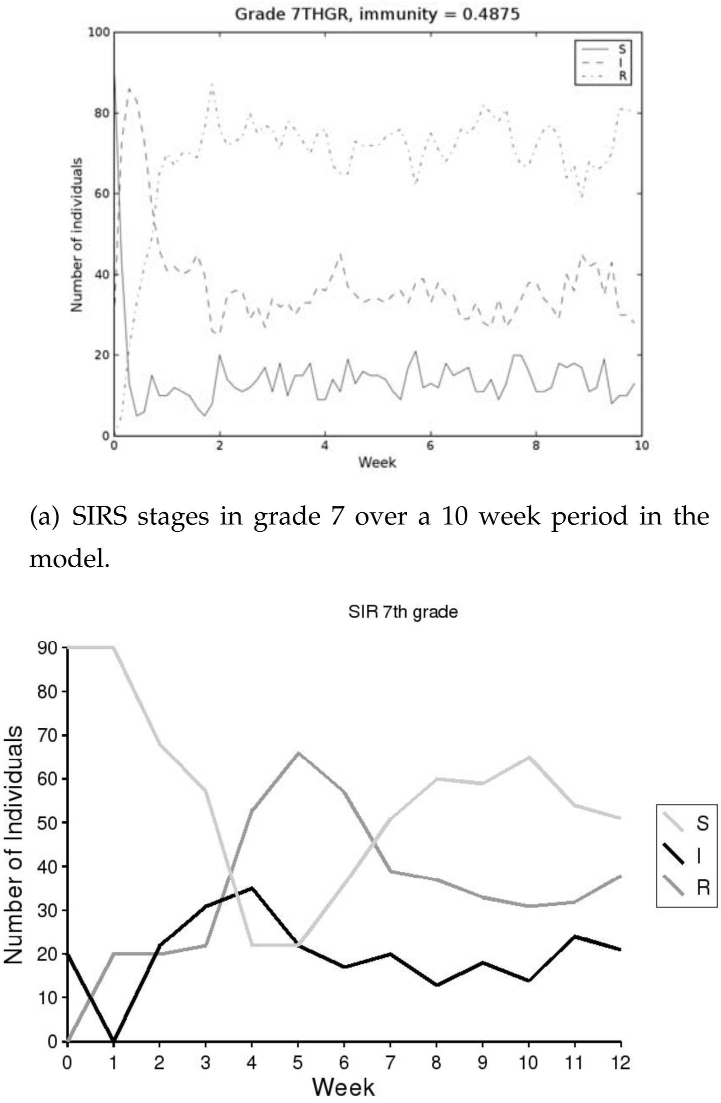
Figure 4.3. Plot of infections for K4 kindergarten classes. Plot of number of people in S, I, and R stages over a 10 week period in the data and model for K4 (kindergarten, age 4) classes. Time is in days for the model, and weeks for the actual data (10 weeks = 70 days total).



(a) SIRS stages in kindergarten 5 over a 10 week period in the model.

(b) Actual data of SIRS stages from kindergarten 5 over a 10 week period.

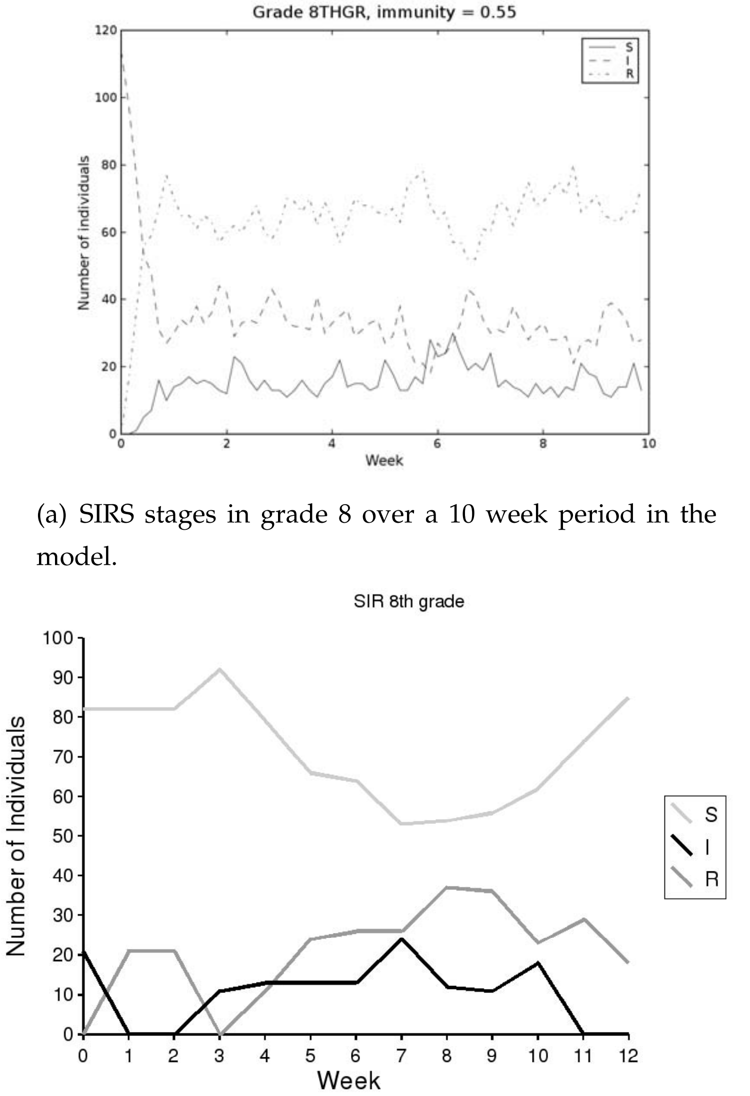
Figure 4.4. Plot of infections for K5 kindergarten classes. Plot of number of people in S, I, and R stages over a 10 week period in the data and model for K5 classes.



(a) SIRS stages in grade 7 over a 10 week period in the model.

(b) Actual data of SIRS stages from grade 7 over a 10 week period.

Figure 4.5. Plot of infections for grade 7 classes. Plot of number of people in S, I, and R stages over a 10 week period in the data and model for grade 7 classes.



(a) SIRS stages in grade 8 over a 10 week period in the model.

(b) Actual data of SIRS stages from grade 8 over a 10 week period.

Figure 4.6. Plot of infections for grade 8 classes. Plot of number of people in S, I, and R stages over a 10 week period in the data and model for grade 8 classes.

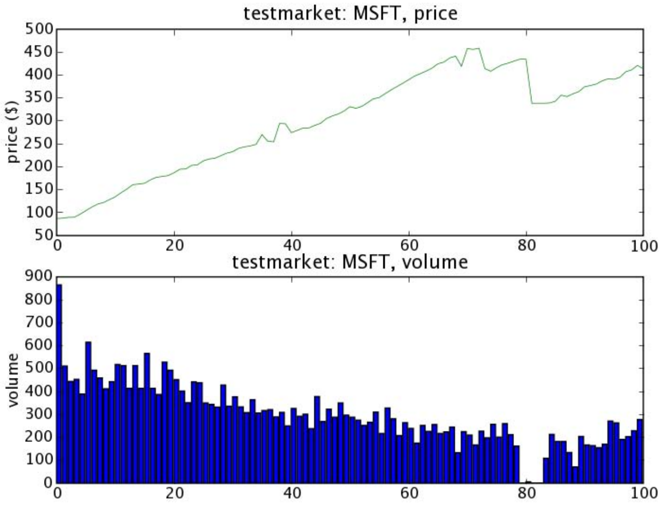


Figure 4.7. Share market graphs showing inflation. Graphs of the share price and volume traded when the initial bank balance is in the range 1c–$10. Observe that inflation occurs, compared with the other graphs that show either flat price trends or varying degrees of booms due to the spread of memes.

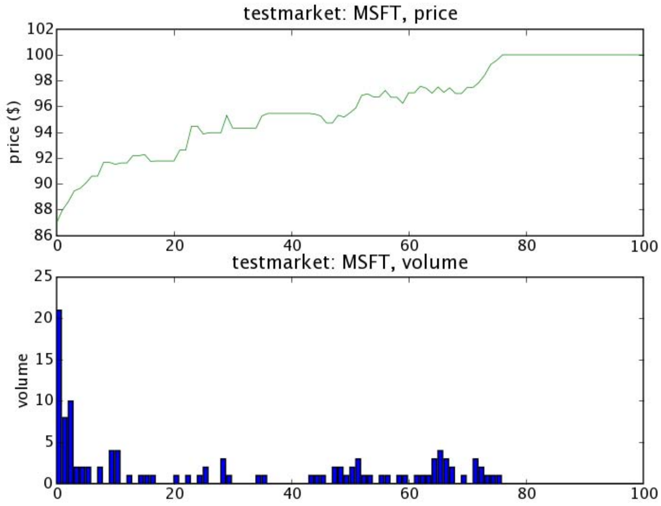


Figure 4.8. Share market graphs for totally disconnected network. Graphs of the share price and volume traded when the initial bank balance is in the range 1c–$10 and the social network is totally disconnected ($p = 0$). Thus no memes spread and there are no booms or busts, merely fluctuations about the initial share price.

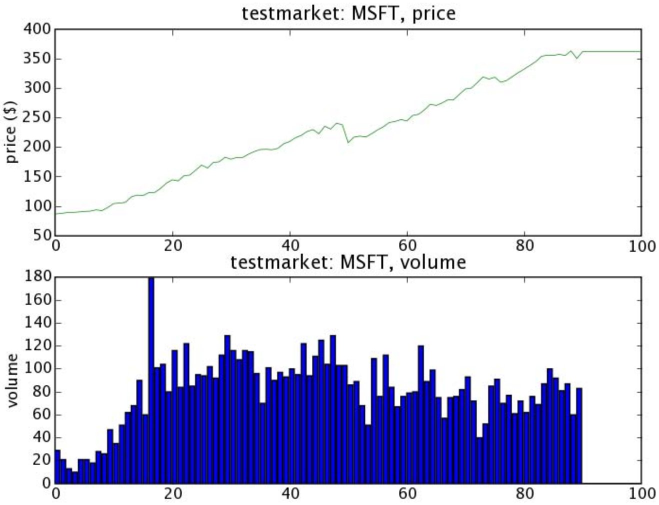


Figure 4.9. Share market graphs for partly connected network. Graphs of the share price and volume traded when the initial bank balance is in the range 1c–$10 and the social network is partly connected ($p = 0.5$). Thus memes can spread and there is a boom occurring, with some fluctuations. At the end there is a period where no trades occur due to non double auctions clearing, and thus the price remains static.

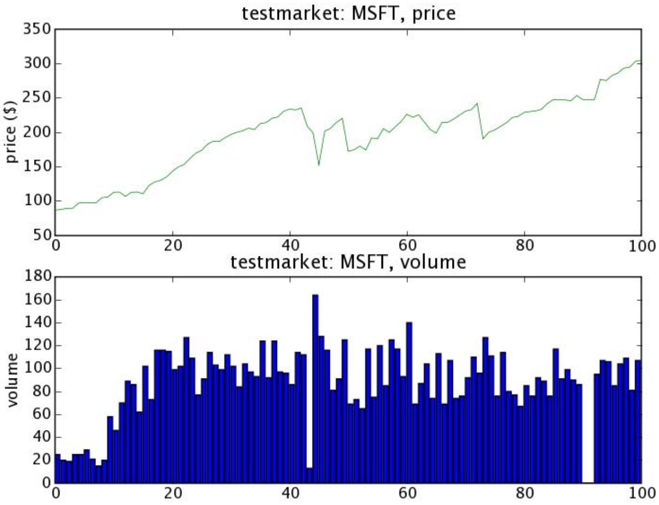


Figure 4.10. Share market graphs for totally connected network. Graphs of the share price and volume traded when the initial bank balance is in the range 1c–$10 and the social network is fully connected ($p = 1$). The social network is large enough that differing sets of memes can start to occur and we see bust cycles in addition to boom cycles.

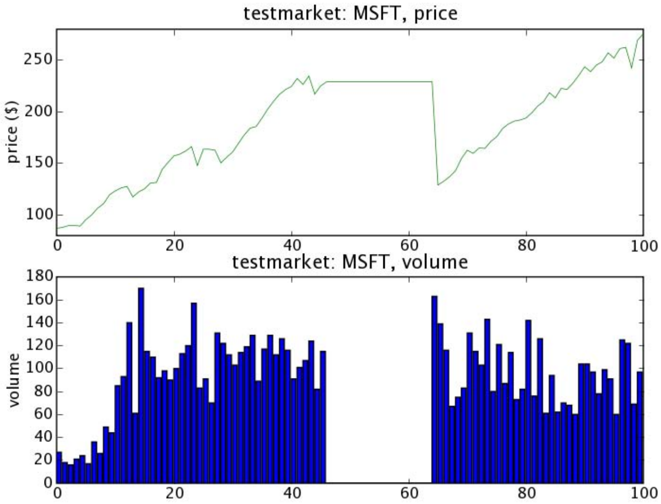
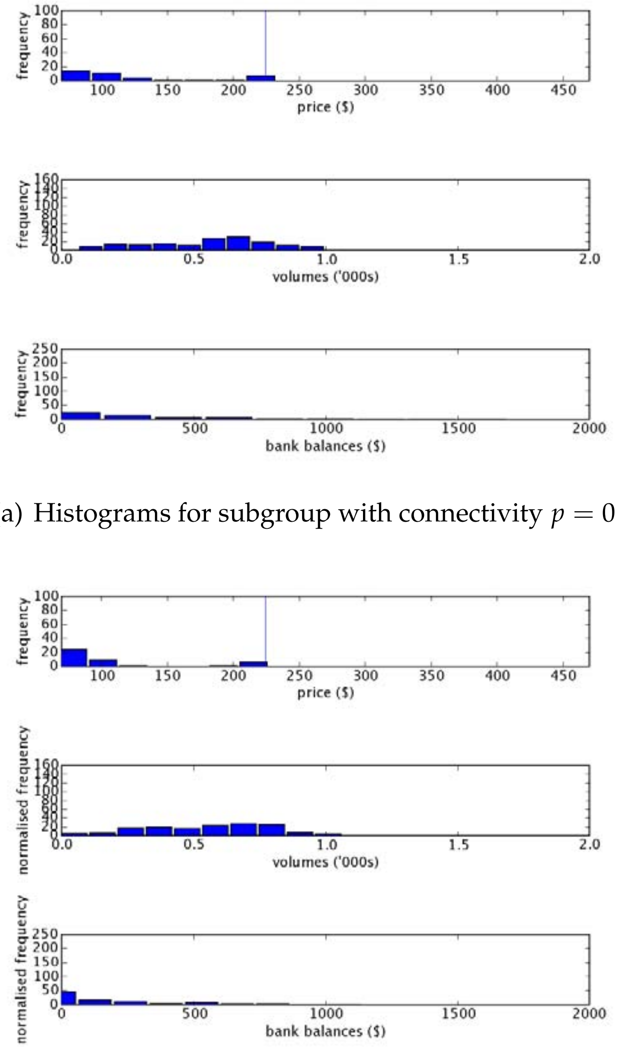


Figure 4.11. Share market graphs for a network with two subgroups of traders. The two subgroups (of 150 traders each) are connected with probability $p = 0.2$ of links from each group to the other. The differences in memes between the subgroups can diverge such that while one group has a buy meme, driving the share price up, the other can initiate selling, driving the share price down to the point where the other meme becomes unsustainable and then a "bust" or crash occurs. One can also see a region where the memes are too diverse, which means in the double auction market that no trades occur. Trading is still occurring in the other listed shares (graphs not shown).



(a) Histograms for subgroup with connectivity $p = 0.5$.

(b) Histograms for subgroup with connectivity $p = 1$.

Figure 4.12. Distribution graphs for a network with two subgroups of traders. This figure shows the histograms of the distribution of meme prices, meme volumes, and bank balances at time step 40 (of 100) for the two subgroups of different connection probabilities. The vertical line in the graphs of the price distributions is the current market price. The two subgroups (of 150 traders each) are connected with probability $p = 0.2$ of links from each group to the other. Here you can see there are two distinct sets of memes forming.

4.5 Conclusions and future work

4.5.1 SIRS model

The model plots show qualitatively similar features to the actual data, in that the viruses quickly reach a steady state from just a few infections at the start of each term, with oscillations as the virus spreads from class to class and back again. This is due to the network structure, which is captured as accurately as possible.

It is difficult to determine causation for the trends between grades at the moment. One would like to show that immunity increases with age. This seems to be a slight trend in the actual data when matched with the current model. The current model does not match accurately enough the initial conditions of class size and number of classes per grade—the effects one can observe could simply be due to this. Furthermore, the structure of the friendship cliques could be changing. One could try and determine this from patterns of infection within a grade, although the data is limited and this fact presents general problems for drawing strong conclusions. The data for grade K4, Figure 4.3(b), and seventh grade, Figure 4.5(b), show an interesting cycling pattern.

Both the frequency data and the SIR data suggest a similar trend in progressing from grade K4 to grade K5 and grade 7 to grade 8 but there is no discernible trend in progressing from grade K4 all the way up to grade 8. Perhaps there is a separate virus that moves through the higher grades that have immunity to the first one. Or perhaps the grades are just simply too "disconnected", or maybe the data too incomplete? Another consideration is simply better hygiene through education; education on prevention strategies has been shown to correlate with transmission of viruses (Vandemoortele and Delamonica 2000). Items for future work include:

1. Adding an incubation state (E) (Earn *et al.* 2000). Rhinoviruses have a 8-12 hour incubation period during which a person is infected but incapable of passing on the virus (Harris and Gwaltney 1996). This is not captured in the SIRS model.

2. Make α_{ij} a function of time. One possible method is to increment α_{ij} if virus i performs worse than the "best" virus j, as measured by total number of infections by virus j.

3. Exploring analytical solutions using the correlation (stochastic differential) equation approach (Rand 1999, Keeling and Rand 2001).

4. Explore the methods developed by Koplik *et al.* (1988) and extend them to the spread of viruses in random networks.

4.5.2 Memes

Clearly this model shows the effects of cash in inflation, and that reducing spending power is an effective way of reducing inflation. Boom effects still occur, however, as the price is governed by memes that have no relation to the underlying dynamics of business but rather to the success of memes. Similar effects are, of course, seen in real life such as the recent IT share boom where the price of the shares bore little-to-no relation to the underlying value, and the continual buying of share meant more people making money, a pattern copied until eventually it was unsustainable and prices crashed. This could be built in to the model, where the value of the meme is a function of how the other person is succeeding as a whole. Other ideas for future work include:

- "Pump and dump" nodes, where one trader actively spreads a meme influencing other traders to buy, in order to sell their own share at a higher price.
- Agents that spread memes to many other traders, yet they do not trade themselves, thus mimicking newspapers and other mass media.
- Algorithms as memes—where rather than the memes being simply information, instead they contain distinct trading strategies (or mixes of multiple strategies).

4.5.3 General conclusions and future work

Different network structures have clear implications for the spread of both viruses and memes, and this is an area that needs further investigation. General statistical methods need to be applied to the data, with more simulation runs performed, in order to verify the effects seen, and also for a more quantitative comparison with real data.

To summarise the contributions of this work they were: the development of models for the spread of viruses in a school and the spread of ideas (about share price) in a sharemarket, and interesting results on the impact of (social) network structure on these systems.

Chapter 5

Drosophila

IN this chapter a cellular automaton for exploring gene interactions in segmentation of *Drosophila* (fruit fly) larvae is presented. Beginning with the expression levels of maternally expressed genes such as *bicoid*, this simple model successfully produces the distinctive expression pattern of the *even-skipped* gene in developing larvae. This work highlights how complex gene interactions in a developing organism can nonetheless be modelled using simple rules.

5.1 Introduction

In this chapter a cellular automaton is presented for exploring the segmentation of *Drosophila melanogaster*, commonly known as the fruit fly. The development of segments is controlled by a number of morphogens, proteins that act to control and regulate the development and shape of an organism (Wolpert *et al.* 1998). Although partial differential equations have been used to explore morphogenesis (Turing 1952, Holloway *et al.* 2003), it can be argued that cellular automata offer a more powerful, flexible approach for capturing the key features of morphogenesis, of which the segmentation of *Drosophila* is one example. To quote John Holland (the inventor of genetic algorithms),

> *Turing (1952) did manage to use PDE's to design a model that started from symmetric initial conditions, but produced an asymmetric variegated pattern, much like the colour pattern of a Holstein cow. Even this simple formulation was mathematically intractable: Turing could observe specific examples of the dynamics, but he could derive no general consequences from the mathematical model. In fact, he depended on a computer-based version of the model to exhibit the dynamics of asymmetric pattern formation. Little has been done mathematically since then, and the problem remains much as it was. (Holland 1995)*

Turing's work on morphogenesis has, however, proven useful, including successes in describing *Drosophila* (Holloway *et al.* 2003). Here an alternative, cellular automaton approach is introduced, which is naturally suited to describing interactions within and between cells. Cellular automata allow for a larger set of rules governing the time-evolution of local protein gradients than PDEs.

5.1.1 Novel contributions

The novel contributions of this work are:

1. Development of a cellular automaton for the modelling of a gene network in *Drosophila*, with reproduction of some of the segmentation pattern that occurs in *Drosophila* larvae.
2. Exploration of the robustness of the gene network in *Drosophila*, in terms of variation of this segmentation with different threshold sensitivities.

5.2 Gene expression in Drosophila

5.2.1 Overview

The set of genes involved in Drosophila form a complex network with both positive and negative feedback and branching and converging pathways across and between levels in a multilevel network (Nijhout 2003). Although the network may appear simple (see Figure 5.1), such simplicity can give rise to highly nonlinear behavior (Nijhout 2002, Holland 1998). Here a subset of genes from the maternal, gap, and pair-rule classes is considered. The extensive research that has been undertaken into their interactions is reviewed in this Section.

NOTE: This figure is included on page 57 of the print copy of the thesis held in the University of Adelaide Library.

Figure 5.1. Network of Drosophila gene interactions. This figure, from Nijhout (2003), shows a network of some of the maternal, gap, pair-rule, and segment polarity class genes. Observe the branching both within and between layers, which gives rise to complex, nonlinear behaviours.

The maternal genes are those that are transcribed in the mother, and the mRNAs are then transported to the oocyte (egg) where they are expressed. These then in turn regulate the expression of the gap genes, which then regulate expression of the pair-rule class of genes. In the following sections, a simple model is built that includes these interactions and show how this leads to the expression of even-skipped stripe two — where the term stripe two refers to the second of seven stripes that appear in the later stages of normal (wild-type) embryo development, in the formation of a segmented body. In the remainder of this Section, key genes involved in this segmentation formation are discussed, and these are all used in the model.

5.2.2 Bicoid

Bicoid is a morphogen translated from maternally expressed mRNA (messenger ribonucleic acid), the first step in determining the anterior-posterior (AP) axis (Houchmandzadeh *et al.* 2002). The expression of *bicoid* is also affected by other maternal effect genes called *exuperantia*, *swallow*, and *staufen* (Fronhöfer and Volhard 1987, Stephenson *et al.* 1988, St. Johnston *et al.* 1989). Localization of *bicoid* mRNA begins during oogenesis, and is controlled by a number of genes including *homeless* (Ray and Schüpbach 1996). As can be seen in Figure 5.2, *bicoid* expression follows an exponential decay curve.

5.2.3 Nanos

Another morphogen translated from maternally expressed mRNA that helps determine the posterior region of the *Drosophila* larva is *nanos* (Wang and Lehmann 1991). Although other maternally expressed genes are involved in setting up the posterior formation, such as *oskar* and *cappucino* (Lehmann and Volhard 1986, Manseau and Schüpbach 1989) , nanos plays a critical role in setting up the posterior region by repressing expression of *hunchback* and *bicoid* (Wang and Lehmann 1991, Lehmann and Volhard 1991). Figure 5.3 shows the expression of *nanos* in a *Drosophila* larva.

5.2.4 Staufen

Another maternally expressed morphogen is staufen (St Johnston *et al.* 1991). This is expressed in the pattern shown in Figure 5.4. Staufen regulates the expression of *hunchback*, although it is unclear if this is through directly regulating the expression of *hunchback* or the expression of *bicoid*.

NOTE: This figure is included on page 59 of the print copy of the thesis held in the University of Adelaide Library.

Figure 5.2. Wildtype expression of bicoid. This figure, from Houchmandzadeh et al. (2002), shows the wildtype (wt) expression of the bicoid protein in a Drosophila larva. The top image shows the expression level using a grayscale intensity. The bottom image shows the numerical values of the intensity as a function of normalised length, determined from the image, and an exponential decay curve fitted to the data. The exponential curve takes the form $I = e^{-\lambda x}$ where I is intensity, x is position, and $\lambda = \surd D/\omega$ for D the diffusion coefficient and ω the protein degradation rate.

NOTE: This figure is included on page 59 of the print copy of the thesis held in the University of Adelaide Library.

Figure 5.3. Maternal expression of nanos mRNA. This figure, from Wang and Lehmann (1991), shows the maternally expressed nanos mRNA in Drosophila,which is highly localised to the posterior region.

NOTE: This figure is included on page 60 of the print copy of the thesis held in the University of Adelaide Library.

Figure 5.4. Wild-type expression of staufen protein. This figure, from St Johnston et al. (1991), shows the wild-type expression of staufen protein in both a freshly-laid larva (A) and a mid-cleavage stage larva (B).

5.2.5 Hunchback

The expression of hunchback is clearly regulated by bicoid, as can be seen in Figure 5.5. This expression is a positive feedback cycle, with both bicoid and hunchback itself driving further up-regulation (higher expression) of hunchback (Houchmandzadeh et al. 2002,Wu et al. 2001). Wu et al. suggest that positive feedback is the only mechanism for the second hunchback stripe in the posterior region. However, Houchmandzadeh et al. show that mutations in staufen affect the boundaries of hunchback by a mechanism other than by staufen changing regulation of bicoid expression. Further, staufen expression is localised to both the poles (the ends of the larvae) (see Figure 5.4). Hunchback expression is also repressed by nanos in the posterior region (Irish et al. 1985), and possibly by knirps (Sauer and J¨ackle 1995).

NOTE: This figure is included on page 61 of the print copy of the thesis held in the University of Adelaide Library.

Figure 5.5. Expression of bicoid and hunchback proteins at different temperatures. This figure, from Houchmandzadeh et al. (2002), shows the levels of bicoid (a) and hunchback (b) as a function of normalised length, for various different environmental temperatures at which the embryos were growing. Note the small spread of hunchback levels for quite a large spread of bicoid levels, especially in the region highlighted in Subfigure (b) where hunchback falls sharply. Subfigures (c)-(f) show all the profiles for the boxed region in Subfigure (b) for temperatures of 9°C, 18°C, 25°C, and 29°C respectively.

5.2.6 Kru¨ppel

Hoch et al. (1991) carried out a detailed study of Kr¨ uppel activation and found that bicoid activates expression of Kr¨ uppel, while hunchback represses it (Hoch et al. 1991).

In other work, they also found that the Kr¨uppel promoter contains binding sites for activation by bicoid repression by knirps (Hoch et al. 1992). The typical pattern of Kr¨uppel expression is shown in Figure 5.6.

NOTE: This figure is included on page 62 of the print copy of the thesis held in the University of Adelaide Library.

Figure 5.6. Expression of Kru¨uppel. This figure, from Small et al. (1992), shows the expression of Kr¨uppel in the darker regions.

5.2.7 Knirps

Knirps expression is activated by bicoid in the anterior end of the Drosophila larva (Rothe et al. 1994). Knirps expression in wild-type Drosophila is shown in Figure 5.7.

NOTE: This figure is included on page 62 of the print copy of the thesis held in the University of Adelaide Library.

Figure 5.7. Time evolution of knirps expression. This figure, from Pankratz et al. (1990), illustrates the expression of knirps in early Drosophila development (a) and at a later stage (b) where the anterior knirps stripe has fully formed.

5.2.8 Giant

Giant is activated by bicoid and repressed by hunchback (Eldon and Pirrotta 1991), and its expression pattern is shown in Figure 5.8.

5.2.9 Even-skipped

Even-skipped expression is controlled by activation by hunchback and bicoid (Small et al. 1992, Frasch and Levine 1987). Knirps can act to repress bicoid-mediated activation by binding to promoter sites near even-skipped sites (Arnosti et al. 1996). Pankratz et al. (1990) detail the importance of Kr¨ uppel and knirps in regulating stripe formation but acknowledge other gap genes may be involved. Small et al. describe the involvement of Kr¨uppel, giant, bicoid, and hunchback in the regulation of even-skipped stripe two (Small et al. 1992). The pattern of even-skipped expression is shown in Figure 5.9.

NOTE: This figure is included on page 63 of the print copy of the thesis held in the University of Adelaide Library.

Figure 5.8. Expression of giant and even-skipped. This figure, from Small et al. (1992), shows the expression of giant in the darker region, and the position of even-skipped stripe two in the narrow darkest region to the left of centre.

NOTE: This figure is included on page 63 of the print copy of the thesis held in the University of Adelaide Library.

Figure 5.9. Expression of even-skipped. This figure, from Small et al. (1992), shows evenskipped expression in Drosophila, with stripe two, regulated by Kr¨uppel, giant, bicoid, and hunchback, shown in a darker colour.

5.3 Cellular automaton modelling

5.3.1 Overview

In cellular automata, discrete locations in space (cells) are updated at (discrete) time t based on the history of their own state and those of some set of the other cells. In the model herein, cells are updated based on their state and the states of neighbouring cells at the previous time step. The type of neighbourhood used is the Moore neighbourhood given in Figure 4.1(b), since this seemed the most realistic. The rules for updating are a set of expression level functions, defining the expression of a protein in terms of the current state—expression of proteins—of the cell and (for some, but not all proteins) the expression levels in the neighbouring cells.

For each of the expression level functions, as much of the biological information was used as possible. Where this is unclear or uncertain, reasonable assumptions were made about the biology and/or information was left out of the model. It is found that in determining the overall position of the stripes, the unused information makes little difference to the general trends when compared with the actual expression levels shown in Figures 5.2 to 5.9. Expression levels are all functions of discrete cell position (x, y, z) and discrete time t. The normalised position along the anterior-posterior axis is denoted by x, where $x = 0$ corresponds to the most anterior position and $x = 1$ corresponds to the posterior. Similarly, y is the normalised position in the ventral-dorsal axis, and z is the normalised position in the medial-lateral axis. Many of the formulae used for computing the change in expression levels are based on simple thresholds. In this chapter it is shown that this can produce results concordant with biological observations, but cannot explain the observed robustness to variations in expression levels of the genes being thresholded, with much trial and error needed to find the settings of the thresholds.

5.3.2 Bicoid model

The maternally expressed pattern of *bicoid* expression is a fixed pattern, with an exponential decay function from the anterior end to the posterior end, and is of the form

$$E_b\left(x, y, z, t\right) = \exp\left(-2|0.05 - x|\right), \tag{5.1}$$

where $E_b\left(x,y,z,t\right)$ is the bicoid level at normalised position $\left(x,y,z\right)$ at time t, $E_b \in \left(0,1\right)$. Note that two exponential decays away from a normalised position of 0.05 are used, to reflect better the true gradient as shown in Figures 5.2 and 5.5.

5.3.3 Nanos model

The maternal *nanos* expression is localised quite specifically and uniformly, in the pattern

$$E_n\left(x,y,z,t\right) = \begin{cases} 0.7, & x > 0.9, \\ 0, & \text{otherwise}, \end{cases} \tag{5.2}$$

where E_n is the level of nanos at position $\left(x,y,z\right)$ at time t.

5.3.4 Staufen model

The maternally expressed pattern of *staufen* is treated as a pair of exponentially decaying functions, starting at both ends, roughly in line with the general trends observed by St Johnston *et al.* (1991) (see Figure 5.4). We can also make the assumption that this has an exponential trend in line with diffusion equations and is similar to bicoid expression. Thus the following function can be used for describing staufen expression,

$$E_s\left(x,y,z,t\right) = \frac{3}{4}\left(\exp\left(-3x\right) + \exp\left(3\left(-1+x\right)\right)\right), \tag{5.3}$$

where $E_s\left(x,y,z,t\right)$ is the level of staufen at position $\left(x,y,z\right)$ at time t, $E_s \in \left(0,1\right)$.

5.3.5 Hunchback model

For a cell at position (x, y, z, t), the level E_h of hunchback is given by the following equation,

$$E_h(x, y, z, t) = \begin{cases} \mathrm{sig}\big(E_h(t-1) + E_b(t-1 & \\ +E_s(t-1) - E_k\,(t-1)\big), & \text{if } \mathrm{sig}\big(E_h(t-1) + E_b(t-1) \\ & +E_s(t-1) - E_k(t-1)\big) > 0.675 \\ & \text{and } E_n(t-1) < 0.1, \\ 0, & \text{otherwise,} \end{cases} \tag{5.4}$$

where

$$\mathrm{sig}(y) = \frac{1}{1+e^{-y}}, \tag{5.5}$$

and $E_h\,(x, y, z, t) \in [0, 1]$ is the level of hunchback at position (x, y, z) at time t, and E_k is defined below for knirps. Note that one can omit the position of the expression levels to save space, as these are all (x, y, z), and thus show that the expression of *hunchback* in a cell depends only on the levels of the other proteins in the cell at the previous time step, $t-1$.

5.3.6 Krüppel model

For a cell at position (x, y, z, t), the level of Krüppel, E_r, is given by the following equation,

$$E_r\,(x, y, z, t) = \begin{cases} 0.7, & 0.5 < E_h(x, y, z, t-1) < 0.85 \\ & \text{and } E_b(x, y, z, t-1) > 0.4 \\ & \text{and } E_k(x, y, z, t-1) < 0.65 \\ 0.1, & \text{otherwise,} \end{cases} \tag{5.6}$$

5.3.7 Knirps model

For a cell at position (x, y, z, t), the level of knirps, E_k, is given by the following equation,

$$E_k\left(x,y,z,t\right) = \begin{cases} \operatorname{sig}\big(E_b(x,y,z,t-1)\big), & E_b(x,y,z,t-1) > 0.8 \\ 0.8, & E_b(x,y,z,t-1) > 0.4 \\ & \text{ and } E_h(x,y,z,t-1) < 0.55, \\ 0.1, & \text{otherwise,} \end{cases} \tag{5.7}$$

5.3.8 Giant model

For a cell at position (x, y, z, t), the level of giant, E_g, is given by the following equation,

$$E_g\left(x,y,z,t\right) = \begin{cases} 0.7, & E_h(,x,y,z,t) > 0.75 \\ & \text{ and } E_k(x,y,z,t-1) < 0.6 \\ & \text{ and } E_r(x,y,z,t-1) < 0.6, \\ 0.1, & \text{otherwise.} \end{cases} \tag{5.8}$$

5.3.9 Even-skipped model

Based on work by Small *et al.* (1992) and Pankratz *et al.* (1990), one can use the following equation for E_e, the level of even-skipped,

$$E_e(x,y,z,t) = \begin{cases} \operatorname{sig}\big(\tfrac{1}{6}E_h(x,y,z,t-1) & \\ +\tfrac{5}{6}E_b(x,y,z,t-1)\big), & \dfrac{1}{|\mathcal{N}|}\displaystyle\sum_{\boldsymbol{p}\in\mathcal{N}} E_r\left(\boldsymbol{p},t\right) > 0.2 \\ & \text{and } \dfrac{1}{|\mathcal{N}|}\displaystyle\sum_{\boldsymbol{p}\in\mathcal{N}} E_g\left(\boldsymbol{p},t\right) > 0.2, \\ 0, & \text{otherwise,} \end{cases} \tag{5.9}$$

where $\mathcal{N}$ is the neighborhood of points
$\mathcal{N} = \{(x', y', z') : |x' - x| \leq 1, |y' - y| \leq 1, |z' - z| \leq 1, (x', y', z') \neq (x, y, z)\}$ about the point (x, y, z).

5.4 Results

5.4.1 Gene expression in the *Drosophila* model

The results herein show the x-y plane in its usual Cartesian arrangement and the choice of co-ordinates as detailed in Subsection 5.3.1 ensures the images produced by the software are in the standard orientation used for displaying *Drosophila* expression levels, with the anterior to the left and the dorsal to the top. The expression levels for each gene lie in $[0, 1]$, these are mapped linearly onto the range for each 8 bit ($\{0, \ldots, 255\}$) RGB (red, green, blue) component, up to a maximum of three genes (one per component). Figure 5.10 shows the simulated bicoid expression levels. They appear quite similar to the actual levels of bicoid expression seen in real *Drosophila* as shown in Figure 5.2, although the exponential tail off towards the anterior end is more clear. Figure 5.11 shows the nanos expression, set to be expressed at the most posterior region, and not expressed elsewhere. With staufen, a fixed expression pattern was also used, and this is shown Figure 5.12.

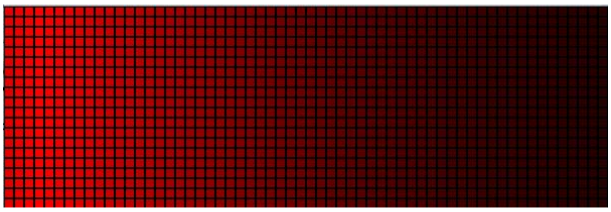


Figure 5.10. Modelled expression of bicoid. This figure shows the level of bicoid in a set of cells representing a cross-section (fixed z) through a *Drosophila* larva, with the anterior to the left and the dorsal to the top. The colour intensity represents the expression level: darker for low levels of expression and lighter for high levels. Note the exponential decay of intensity as we move away from a normalised x position of 0.05, which is three cells from the left (anterior) side.

Figure 5.13, shows the level of hunchback after its expression has stabilised into a fixed pattern, and Figure 5.14 shows both hunchback and bicoid simultaneously for the same point in time. Note that hunchback has a well defined boundary in the middle region, whereas bicoid has a continuous gradient. Experimentation (not shown here) revealed that the position of this boundary varied little with changes in the bicoid gradient. This suggests that the proposals, by Houchmandzadeh *et al.* regarding the effect of other genes including positive feedback from hunchback itself in Houchmandzadeh *et al.* (2002), are correct.

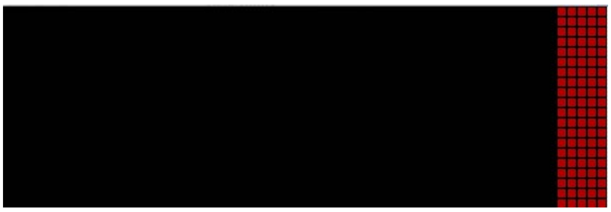


Figure 5.11. Modelled expression of nanos. The level of nanos in a set of cells representing a cross-section (fixed z) through a *Drosophila* larva is shown. The red band on the right indicates high levels of expression and the darker region represents low expression levels.

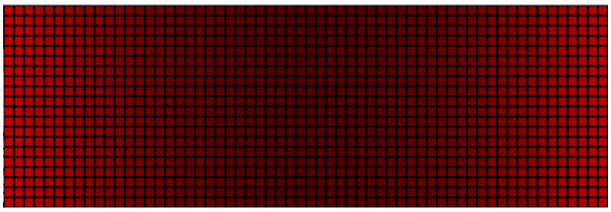


Figure 5.12. Modelled expression of staufen. The level of staufen in a set of cells representing a cross-section (fixed z) through a *Drosophila* larva is shown in this figure. The colour intensity represents the expression level: darker for low levels of expression and brighter for high levels.

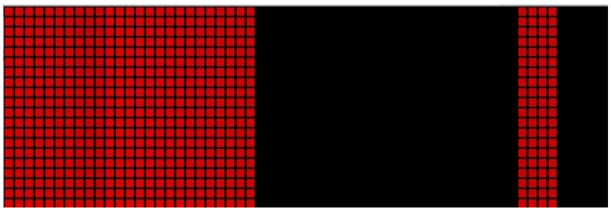


Figure 5.13. Modelled expression of hunchback. This figure shows the level of hunchback in a set of cells representing a cross-section (fixed z) through a *Drosophila* larva. The colour intensity represents the expression level: darker for low levels of expression and lighter for high levels.

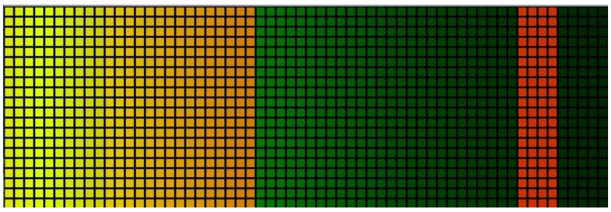


Figure 5.14. Modelled expression of hunchback and bicoid. This figure shows the expression level of both hunchback and bicoid in a set of cells representing a cross-section (fixed z) through a *Drosophila* larva. The pattern is simply an overlay of Figures 5.10 and 5.13, with the lightest regions corresponding to regions where hunchback is expressed.

Figures 5.15, 5.16, and 5.17 show the expression levels for the gap class genes *Krüppel*, *knirps*, and *giant*. These correspond well with the expression levels shown in Figures 5.6, 5.7, and 5.8, even though the set of interactions has been greatly simplified along with the form these interactions take. The simplification has resulted in these stripes being very sensitive to minor changes in bicoid and hunchback expression, with variations in the normalised expression levels of 0.05 in hunchback and bicoid resulting in the absence of these stripes in some cases. This highlights the fact that context of other genes as shown in the network in Figure 5.1 is important in adding robustness to the gape gene expression against variations in maternal gene expression. Robustness could also be gained by mechanisms other than the simple thresholding used (Houchmandzadeh *et al.* 2002, Holloway *et al.* 2003).

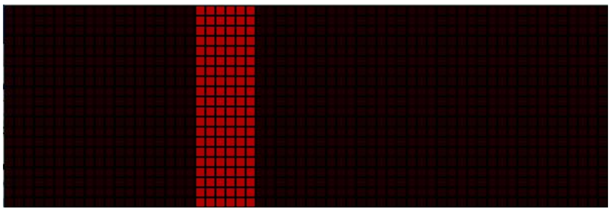


Figure 5.15. Modelled expression of Krüppel. The level of Krüppel in a set of cells representing a cross-section (fixed z) through a *Drosophila* larva. The colour intensity represents the expression level: darker for low levels of expression and lighter for high levels.

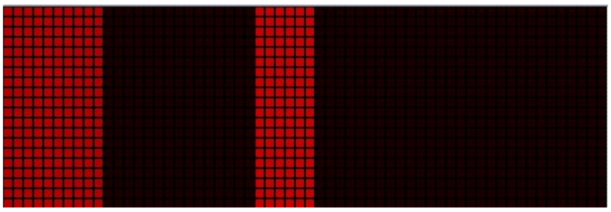


Figure 5.16. Modelled expression of knirps. This figure shows the expression level of knirps in a set of cells representing a cross-section (fixed z) through a *Drosophila* larva. The colour intensity represents the expression level: darker for low levels of expression and lighter for high levels.

Figure 5.18 shows the expression of even-skipped stripe two in the virtual *Drosophila* larva.

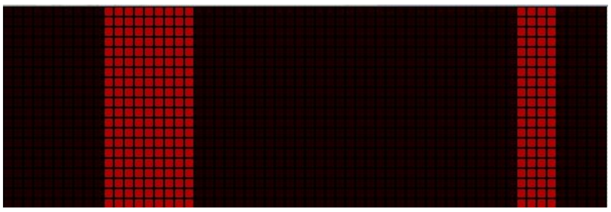


Figure 5.17. Modelled expression of giant. The level of giant in a set of cells representing a cross-section (fixed z) through a *Drosophila* larva. The colour intensity represents the expression level: darker for low levels of expression and lighter for high levels.

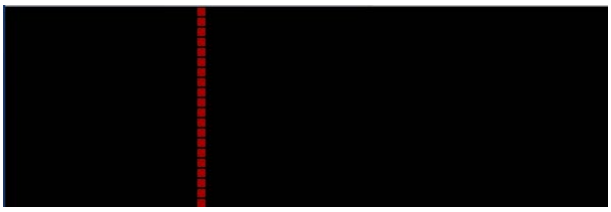


Figure 5.18. Modelled expression of even-skipped. This figure shows the expression level of even-skipped in a set of cells representing a cross-section (fixed z) through a *Drosophila* larva. The colour intensity represents the expression level: darker for low levels of expression and lighter for high levels.

5.5 Conclusions

This chapter has presented a cellular automaton with a simplified set of genes and mostly simple rules governing interaction between those genes. Despite this simplicity, the cellular automaton is able to generate realistic patterns of stripes, up to the even-skipped stripe two. This suggests that *Drosophila* could be modelled quite accurately using a simple yet more powerful model taking into account the other gene interactions, and using interactions consisting of more than thresholding. This would give added robustness to fluctuations in expression in genes higher in the hierarchy. The results from the model indicate that further work is needed to refine the mechanisms by which the gene promoters are acting, to give further clues as to how to best model the interactions.

To summarise the novel contributions of this work: a cellular automaton that models the *Drosophila* gene network was developed that generates some of the known segments in *Drosophila* larvae, and explores the robustness of this gene network.

Chapter 6

The *p53* gene

It is known that *p53* is an important gene, involved in apoptosis (programmed cell death), DNA repair, and cell cycle progression. In this chapter *p53* is considered from two angles: firstly as part of a gene network, with external, environmental inputs into a complex network of interactions between proteins and between proteins and genes, and secondly the effect of mutations on *p53* in a heterogeneous population of tumour cells is explored.

Gene networks are composed of many different interacting genes and gene products (RNAs and proteins). They can be thought of as switching regions in an n-dimensional space or as mass-balanced signalling networks. Both approaches allow for describing gene networks with the limited quantitative or even qualitative data available. This chapter shows how these approaches can be used in modelling a gene network involved in apoptosis (programmed cell death) and DNA repair.

The selective advantages and disadvantages of mutations in the *p53* gene on tumour cells and the heterogeneity of tumour cell populations are explored. Based on an evolutionary computational approach, the model developed considers changes in mutation rate caused by lack of DNA repair processes, and the lack of apoptosis caused by mutations in *p53*. In this chapter it is found that the degree of robustness of *p53* to mutations has a significant effect on the tumour heterogeneity and "fitness", with clinical consequences for people who inherit *p53* mutations.

6.1 Introduction

One in three people are affected by cancer in their lifetime (Caspari 2000). Cancer is caused by multiple DNA mutations that allow cells to proliferate without limit (Gold and Sokolowski 2004). The *p53* protein is a potent tumour suppressor, and plays a major role in maintaining a healthy cell cycle. More information on the cell cycle can be found in Appendix A. During the cell cycle, however, DNA inside cells experiences spontaneous or environmentally induced mutations. Mutations occurring in vital parts of the DNA sequence can be deleterious for the cell.

One type of mutation that can alter DNA sequences is base pair substitution mutations (Montelone 1998). As the name suggests, base pair substitution mutations occur when one of the nucleotide bases is changed; this change is classified as either a transition or a transversion. A transition occurs when a purine (A or G) is changed to another purine or a pyrimidine (C or T, or U in RNA) to another pyrimidine. For example, consider a base pairing of guanine (a purine) and cytosine (a pyrimidine). If during translation the guanine becomes adenine (another purine), the result is the base pairing of adenine and cytosine in the new strand of DNA. If instead, the guanine had been changed to cytosine or thymine (or uracil in RNA), the result would be a base pairing of two pyrimidine's, which is genetically unstable. This type of mutation, when a purine is changed to a pyrimidine or a pyrimidine to a purine, is known as a transversion (Elliott and Elliott 1997).

Because of the speed at which transcription occurs, there is a large window of opportunity for errors to occur. Some mutations, however, can be silent,

1. if the mutation results in a change of the codon, but not the amino acid it codes for, or

2. if it alters the amino acid, but not the protein.

Conversely, mis-sense mutations alter the genes such that the amino acid coded for is changed, or codons that originally coded for amino acids are transformed into stop codons and vice-versa (Montelone 1998). Mutations to either the *p53* gene itself, or the pathways activated by *p53* are the single most common mutations found in cancer and have been implicated in over 50 per cent of cancer cases (Janus *et al.* 1999, Zhou and Elledge 2000). Mutation of the *p53* gene or its effector pathways results in malignant tumours (Brodeur and Lowe 1999).

In healthy cells, the *p53* protein is present at very low concentrations, in an inactive form. Levels of *p53* in the cell are maintained by a negative feedback loop between it and another protein Mdm2, (Sherr 1998, Vogelstein *et al.* 2000). There are several checkpoints in the cell cycle which can prevent the cell from progressing through the cell cycle if DNA damage is detected. Activation of the *p53* protein can occur at two of these checkpoints in the cell cycle. If DNA damage is detected, *p53* is indirectly activated and stops cells in the G1 and G2 phase from progressing, slows down cells in the S phase and induces transcription of repair genes (Elledge 1996). For more information on the cell cycle, refer to Appendix A.

Forms of DNA damage shown to activate *p53* include those caused by ultraviolet (UV) light, other ionizing radiation, and exposure to radio-mimetic drugs (Lakin and Jackson 1999). Activation of *p53* in response to DNA damage increases the ability of the *p53* protein to bind to DNA (Lakin and Jackson 1999) and causes a decrease in its affinity for Mdm2 and hence a decrease in degradation of itself (Alberts *et al.* 2002). The *p53* protein can induce apoptosis[1] and can also initiate transcription of proteins involved in DNA repair and cell cycle arrest (Zhou and Elledge 2000). Cell cycle arrest restricts the cell from entering into the next phase in the cell cycle, until the DNA damage has been repaired (Alberts *et al.* 2002). If the DNA damage is too severe, then the cell will undergo programmed cell death, known as apoptosis. This can be mediated either by *p53* initiating transcription of proteins such as the Bax protein or directly, by stimulation of the mitochondria to produce excess toxic reactive oxygen species (Vogelstein *et al.* 2000). Table 6.1 identifies some of the mutations that can occur and their affects on the activities of *p53*.

Repair mechanisms play a major role in maintaining the integrity of the genome (Elliott and Elliott 1997). In cells, DNA repair is constantly occurring and is essential for their survival. DNA damage reversal is the simplest form of repair; enzymatic action by DNA ligase repairs simple breaks in one strand of the DNA (Montelone 1998). Another form of DNA repair is damage removal (Wood *et al.* 2001). Although not truly a repair mechanism, cells can also exhibit ways of coping with the damage (Montelone 1998).

Evidence, from both mouse and human models, has shown direct or indirect involvement of *p53* in nucleotide excision repair (Yuan *et al.* 1995, Goukassian *et al.* 2000). Lesions that distort the double helix, such as a T-dimer, are repaired by NER (Elliott

[1] For definitions of key terms, refer to the thesis conventions and glossary in the front matter and Appendix A.

Table 6.1. Effect of mutations on p53. From Vogelstein et al. (2000), this table shows the effect of mutations on p53 and the many ways it may malfunction in tumours.

NOTE: This table is included on page 76 of the print copy of the thesis held in the University of Adelaide Library.

and Elliott 1997). NER involves the excision of an oligonucleotide—this comprises the breaking of phosphodiester bonds, on the same strand, on either side of the lesion—the subsequent gap is filled with the aid of DNA ligase during repair synthesis (Griffths et al. 1996).

When p53 function is reduced due to mutations, the DNA repair capabilities of the cell are reduced (Gold and Sokolowski 2004). The cell is unable to cease cell cycle progression at DNA checkpoints and perform DNA repair, so the cell will continue through the cell cycle and the daughter cells will receive mutated DNA, or incomplete or broken sets of chromosomes (Alberts et al. 2002). Ordinarily, DNA damage, or telomere malfunction would trigger apoptosis through p53 to remove these cells (Lowe and Lin 2000), however, mutations to p53 or its effectors can prevent programmed cell death occurring. The lack of apoptosis causes uncontrolled cell proliferation despite DNA damage and has been shown to promote oncogenic transformations and tumour development in mouse model systems (Gold and Sokolowski 2004, Lowe and

Lin 2000). Attardi and Jacks (1999) performed studies on both homozygous and heterozygous mice for a deletion in the *p53* gene. Their results showed that both groups of mice developed tumours at a high frequency; however the homozygotes had a significantly shorter average time to develop tumours and a shorter life span. Their studies of differences between the two sets of mice also indicated that the affects of *p53* inhibition vary in different tissue types.

The revelation that checkpoint function has been strongly implicated in the prevention of cancer (Elledge 1996), has provided significant breakthroughs in research to developing alternative cancer treatments. Such treatments include the reintroduction of wild type *p53* to *p53* mutant tumour lines, which has been shown to induce apoptosis and tumour regression, when performed in conjunction with chemotherapy (Lowe and Lin 2000). Other suggested therapeutic strategies involving *p53* include the restoration of *p53* function to mutant *p53* tumours by *p53* gene therapy; also, drugs that are able to mimic the effects of *p53* or modify its effector pathways may be useful (Brodeur and Lowe 1999). Detection of *p53* mutations has proven to be a useful diagnostic tool. Polymerase chain reaction (PCR)-based techniques have been used to detect mutant *p53* in exfoliated cells in bladder and lung cancers (Brodeur and Lowe 1999). This would be a useful strategy in early detection of cancers, or identifying potential cancer patients.

While cancer is understood in very broad terms (Spencer *et al.* 2004, Hanahan and Weinberg 2000) and many of the key proteins involved are very well understood (Futreal *et al.* 2004, Arends 2000, Knudson 2002), more research is needed at the gene network level to understand interactions between genes, proteins, and the environment. It is well known that *p53* is an important protein involved in a number of key cell processes (Vogelstein *et al.* 2000), which if disturbed can lead to cancer (Spencer *et al.* 2004, Hanahan and Weinberg 2000). By modelling the *p53* gene network, one can gain a better understanding of its interactions, with implications for cancer treatment.

6.1.1 Novel contributions

The novel contributions of this work are:

1. Casting the p53 gene network of human cells into the mathematical (switching network) framework proposed by de Jong *et al.* (2004b).
2. Exploration of the effect of UV-specific and general DNA damage to cells on the interactions in this p53 network, and the end result effected.

The novel contributions made in collaboration with Melissa Ryan were:

1. Exploration of the rate of mutation acquisition in a model p53 gene, as its DNA repair functions are progressively disabled.
2. Exploration of the effects of p53 mutations on the time to onset of cancer.

6.2 Gene networks

6.2.1 Background

Modelling gene networks gives us a broad, yet important view of vital cell regulatory functions (de Jong *et al.* 2004a, Kauffman 1993). Two techniques that have been developed are mass-balanced signalling networks (Meza *et al.* 2004, Kurata *et al.* 2001, Kærn *et al.* 2003)—of which the Michaelis-Menten kinetics are a particular form (Murray 2002)—and piecewise-linear differential equations (DEs) (de Jong *et al.* 2004b, de Jong *et al.* 2004a, Edwards and Glass 2000), a generalisation on the work in Boolean NK switching networks (Glass 1975, Edwards 2000, Edwards and Glass 2000). Although enzyme kinematics are a useful tool (Murray 2002, Meza *et al.* 2004), the ODEs involved are typically quite stiff which increases the computation time (Kurata *et al.* 2001). Also, the piecewise-linear DE approach gives a better picture of the qualitative nature of the behaviour (de Jong *et al.* 2004a), since they are quick and accurate to integrate (Edwards and Glass 2000) and effective analytical tools exist for their qualitative behaviour (de Jong *et al.* 2004b).

6.2.2 Switching networks

Gene networks present several problems for any sort of quantitative analysis (de Jong *et al.* 2004b, de Jong *et al.* 2004a). Many of the interacting proteins may be unknown, and even between the known interacting proteins, some of the interactions occur only under unknown conditions (de Jong *et al.* 2004a). Furthermore, although techniques such as microarrays and reverse transcriptase-polymerase chain reaction (RT-PCR) give gene expression levels (Shoemaker *et al.* 2001), there is a high variance and measurement error, and they do not typically provide enough detail about the complex interactions to establish any strict rules of behaviour in the gene networks (Moreau *et* *al.* 2002). Even in well-established gene networks, such as the *p53* network, the details of many interactions are still being discovered (Vogelstein *et al.* 2000). Some of the key ones, such as the feedback loop that exists between *p53* and another protein, Mdm2, are lacking clear quantitative values (Vogelstein *et al.* 2000). Consequently, one is left with an incomplete picture of protein interactions, with imprecise, qualitative rules—for example, protein A binds to protein B, or protein C activates transcription of protein D. Basing my analysis on logical switching systems is useful, because:

> logical switching systems *capture major features of a homologous class of nonlinear dynamical systems governed by sigmoidal functions because such systems tend to* sharpen *their responses to* extremal values of the variables*. (Kauffman 1993)*

If one considers a typical sigmoidal function,

$$y = \frac{x^n}{\theta^n + x^n}, \tag{6.1}$$

then it is trivial to show that

$$\lim_{n\to\infty} y = \lim_{n\to\infty} \frac{\frac{x^n}{\theta^n}}{1+\frac{x^n}{\theta^n}} = \begin{cases} 1, & x > \theta \\ 0, & x < \theta, \end{cases} \tag{6.2}$$

where $\theta \neq 0$. Thus in the extreme case, as $n \to \infty$, we have a simple switching, or step function. Piecewise-linear (PL) differential equation (DE) models, originally developed by Glass (1975) and Edwards (2000), offer a way around the problem of a lack of clear quantitative values, in that they can successfully model systems in a qualitative sense, including quantitative data where available. Filippov successfully introduced step-functions, which are important for modelling the switch-like regulatory interactions in gene networks (Filippov 1988).

6.2.3 Methods

As described by de Jong *et al.* (2004b), the rates of change of proteins can be written as

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x})\,\boldsymbol{x} \tag{6.3}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, $\boldsymbol{x} \in \mathbb{R}^n_{\geq 0}$, the set of vectors with n non-negative elements, is the vector containing expression levels (copy number divided by cell volume) of the n proteins, $\boldsymbol{f} = (f_1, \ldots, f_n)'$ consists of the rates of synthesis of each proteins, and $\boldsymbol{g}(\boldsymbol{x}) = \operatorname{diag}(g_1, \ldots, g_n)$ is the rate of degradation (de Jong *et al.* 2004b). The rate of synthesis of the protein i, $f_i : \mathbb{R}_{\geq 0} \to \{0, 1\}$ depends on the levels of the proteins, $\boldsymbol{x}$ in some fashion,

$$f_i(\boldsymbol{x}) = \sum_{l \in \mathcal{L}} \kappa_{il} b_{il}(\boldsymbol{x}), \tag{6.4}$$

for rate parameter $\kappa_{il} > 0$, and $b_{il} : \mathbb{R}^n_{\geq 0} \to \{0, 1\}$ a regulation function, and $\mathcal{L}$ a possibly empty set of indices of *regulation functions*. The g_i are defined similarly to the f_i, except it is required that $g_i > 0 \forall i$. The *regulation functions* b_{il} are defined in terms of step functions $\mathrm{s}^+, \mathrm{s}^- : \mathbb{R}^2 \to \{0, 1\}$ defined as

$$\mathrm{s}^+(x_j, \theta_j) = \begin{cases} 1, & x_j > \theta_j \\ 0, & x_j < \theta_j \end{cases} \tag{6.5}$$

$$\mathrm{s}^-(x_j, \theta_j) = 1 - \mathrm{s}^+(x_j, \theta_j), \tag{6.6}$$

where $\theta_j \in \mathbb{R}_{\geq 0}$.

A generalised view of the *p53* network, as shown in Figure 6.1, is considered here. Based on Figure 6.1, one can abstract over the proteins apart from *p53* and Mdm2, and instead consider $\boldsymbol{x} = (x_d, x_u, x_p, x_m, x_A, x_C, x_D)'$ where x_d is the level of DNA damage kinases, x_u the level of UV stress kinases, x_p the expression level of *p53*, x_m the level of Mdm2, x_D the combined level of apoptosis (programmed cell death) proteins, x_A for the angiogenesis proteins, and x_C for the cell cycle proteins. The network of interactions is shown in Figure 6.2. The following set of equations is used to model the *p53* network:

$$x_p = s_{pd}\,\mathrm{s}^+(x_d, \theta_d) + s_{pu}\,\mathrm{s}^+(x_u, \theta_u) + s_{pm}\,\mathrm{s}^+(x_m, \theta_m)\,\mathrm{s}^-(x_p, \theta_p) - g_p x_p \tag{6.7}$$

$$x_m = s_m\,\mathrm{s}^+(x_m, \theta_m)\,\mathrm{s}^-(x_p, \theta_p) - g_m x_m \tag{6.8}$$

$$x_A = s_A\,\mathrm{s}^+(x_p, \theta_A) - g_A x_A \tag{6.9}$$

$$x_C = s_C\,\mathrm{s}^+(x_p, \theta_C) - g_C x_C \tag{6.10}$$

$$x_D = s_D\,\mathrm{s}^+(x_p, \theta_D) - g_D x_D, \tag{6.11}$$

where θ_a for some protein (or protein group) a is the level at which the protein b in the term $s_{cb}\,\mathrm{s}^{\pm}(x_c, \theta_a)$ signals for the synthesis of the protein c at rate s_{cb}. No equations for x_d and x_u are given as these are input variables. The inequality $\theta_C < \theta_A < \theta_D$ is used, which has the interpretation that as a first response the cell stops the cell cycle and signals for angiogenisis, and with the highest levels of damage and stress undergoes apoptosis (programmed cell death).

NOTE: This figure is included on page 81 of the print copy of the thesis held in the University of Adelaide Library.

Figure 6.1. *p53*-Mdm2 feedback loop. This figure, by Nakade (2004), shows some of the key players in the p53 network, including the *p53*-Mdm2 feedback loop. DNA breaks and UV stress cause changes in such kinase proteins as ATM kinase and Casein kinase II. A kinase is a protein that adds a phosphor atom to a protein. In this case, a phosphor atom is added to *p53*. *p53* is then termed phosphorylated, and this affects its binding to Mdm2, altering the feedback loop. The increased levels of *p53* then can signal for inhibition of angiogenesis, cell cycle arrest (to allow DNA repair to occur), and/or apoptosis. The actual response, simply halting or undergoing apoptosis, is a function of the input.

For each output variable, there is a target equilibrium position, which exists in some regulatory domain. Switching domains exist on the boundaries between regulatory domains. This can be seen in the representation of the domains for *p53* and Mdm2 in Figure 6.3.

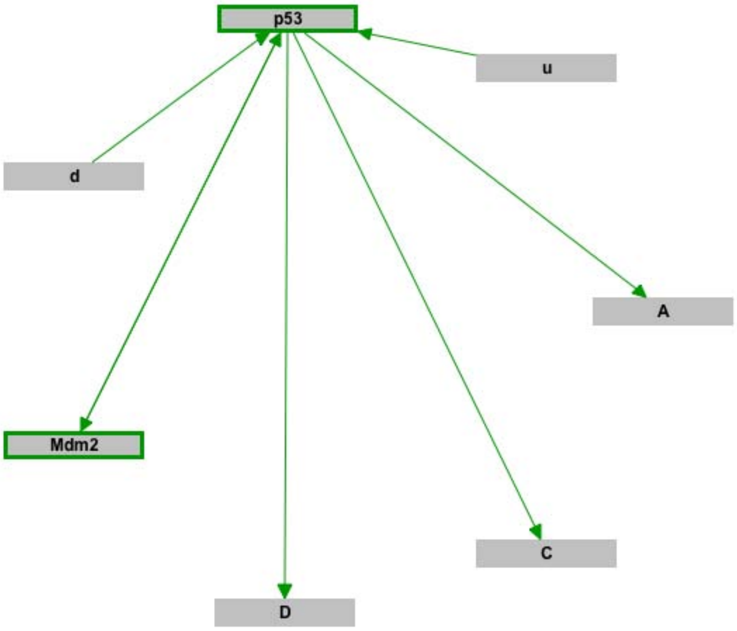


Figure 6.2. Influences on *p53*-Mdm2 network. This network shows with directed arrows the influences of various inputs and proteins on other proteins. UV stress (denoted by u) and DNA damage (d) signal for production of *p53*, which is in a feedback loop with the Mdm2 protein. Downstream targets of *p53* include proteins that signal for angiogeneis (A), halting of the cell cycle (C), and apoptosis (programmed cell death, denoted by D).

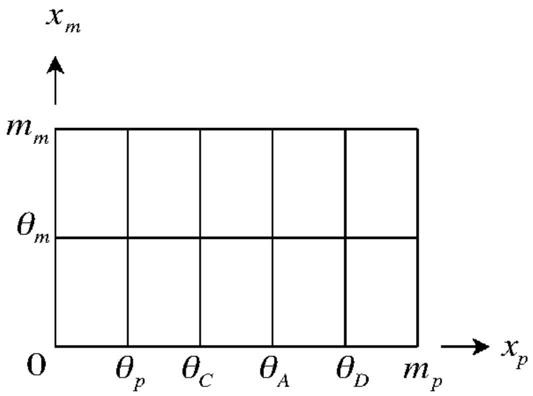


Figure 6.3. Diagram of protein expression values and regulatory domains. The set of protein expression values for $0 \le x_p \le m_p$ and $0 \le x_m \le m_m$ can be drawn as a 2D plane, with points lying either in switching domains, the lines at critical values (θ's), and the regulatory domains (lying between critical values). As the level of *p53* (x_p) increases, it shifts in to different signalling domains. For example if x_p lies in the range (θ_C, θ_A) it means that cell-cycle halting is being signaled for, but not angiogenesis or apoptosis.

The equilibrium positions and the domains are shown in Table 6.2. Note that m_a is the maximum value of expression of protein a, denoted by x_a. For the output protein levels of interest, one defines critical values—if the protein level x_A lies above the critical value c_A, this has the interpretation that protein group A is activated, that is the cell would then be signalling for angiogenesis. These are the primary domains used for the analysis. An alternate set is used to see what happens if the level of *p53* produced in response to various stress levels is reduce. This set is shown in Table 6.3.

There also needs to be a set of initial conditions. Because this is a qualitative model, it is sufficient to simply specify regulatory domains for the initial conditions. Of interest are four different cases,

1. $(x_d, x_u) \in [0, \theta_d) \times [0, \theta_u) = \{(x_d, x_u) \in \mathbb{R}^2 | 0 \leq x_d < \theta_d,\ 0 \leq x_u < \theta_u\}$,
2. $(x_d, x_u) \in [0, \theta_d) \times (\theta_u, m_u]$,
3. $(x_d, x_u) \in (\theta_d, m_d] \times [0, \theta_u]$, and
4. $(x_d, x_u) \in (\theta_d, m_d] \times (\theta_u, m_u]$,

corresponding to varying stresses on the cell; the first one corresponding to no DNA damage or UV stress. In addition, the initial conditions for the other variables must also be specified. These are kept constant as shown in Table 6.4.

Table 6.2. Protein equilibrium positions. The equilibrium position(s) for each output protein level, and the regulatory domains that have been specified for these to occur in. There are several *p53* equilibrium positions—one for each combination of synthesis/binding rates, the others have only one synthesis rate so only one equilibrium position occurs, according to the theorems in de Jong *et al.* (2004b).

Variable	Equilibrium position	Regulatory Domain
x_p	$\left(s_{pd} + s_{pu}\right) / g_p$	$(\theta_D, m_m]$
x_p	$\left(s_{pd} + s_{pu} + s_{pm}\right) / g_p$	$(\theta_D, m_m]$
x_p	s_{pd} / g_p	(θ_A, θ_D)
x_p	$\left(s_{pd} + s_{pm}\right) / g_p$	$(\theta_D, m_m]$
x_p	s_{pu} / g_p	(θ_A, θ_D)
x_p	$\left(s_{pu} + s_{pm}\right) / g_p$	$(\theta_D, m_m]$
x_p	s_{pm} / g_p	(θ_A, θ_D)
x_m	s_m / g_m	$(\theta_m, m_m]$
x_C	s_C / g_C	$(c_C, m_C]$
x_A	s_A / g_A	$(c_A, m_A]$
x_D	s_D / g_D	$(c_D, m_D]$

Table 6.3. Lowered protein equilibrium positions. Similar to Table 6.2, the lowered equilibrium position(s) for each output protein level are listed, and the regulatory domains that have been specified for these to occur in. As discussed previously, *p53* has several equilibrium positions—one for each combination of synthesis/binding rates, the others have only one synthesis rate so only one equilibrium position occurs, according to the theorems in de Jong *et al.* (2004b).

Variable	Equilibrium position	Regulatory Domain
x_p	$\left(s_{pd} + s_{pu}\right)/g_p$	(θ_A, θ_D)
x_p	$\left(s_{pd} + s_{pu} + s_{pm}\right)/g_p$	$(\theta_D, m_m]$
x_p	s_{pd}/g_p	(θ_C, θ_A)
x_p	$\left(s_{pd} + s_{pm}\right)/g_p$	(θ_A, θ_D)
x_p	s_{pu}/g_p	(θ_C, θ_A)
x_p	$\left(s_{pu} + s_{pm}\right)/g_p$	(θ_A, θ_D)
x_p	s_{pm}/g_p	(θ_C, θ_A)
x_m	s_m/g_m	$(\theta_m, m_m]$
x_C	s_C/g_C	$(c_C, m_C]$
x_A	s_A/g_A	$(c_A, m_A]$
x_D	s_D/g_D	$(c_D, m_D]$

Table 6.4. Initial regulatory domains. Initial regulatory domains for *p53* (x_p), Mdm2 (x_m), and output signals for the halting of cell cycle (x_C), angiogenesis (x_A), and apoptosis (x_D).

Variable	Initial domain
x_p	$[0, \theta_A]$
x_m	$[0, m_m]$
x_C	$[0, c_C]$
x_A	$[0, c_A]$
x_D	$[0, c_D]$

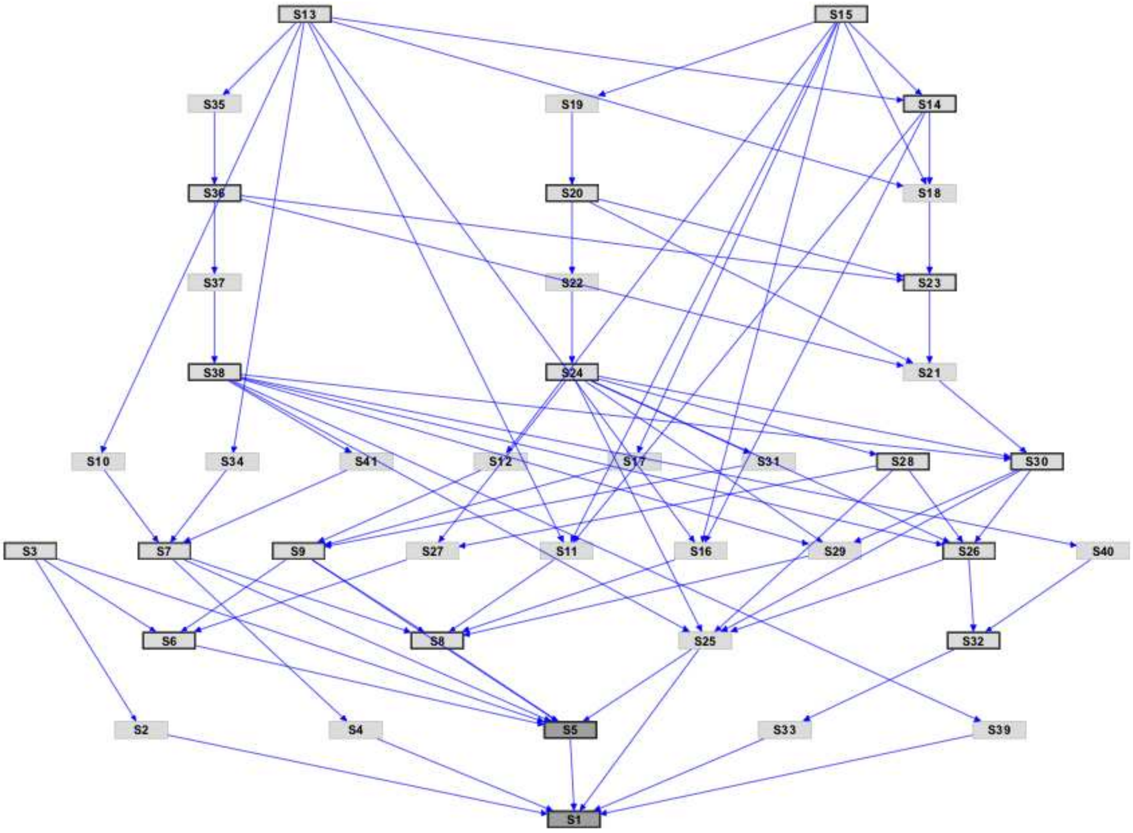


Figure 6.4. Graph of states for no external stress on the cell. This graph shows the final set of states for the input set one, $(x_d, x_u) \in [0, \theta_d) \times [0, \theta_u)$, representing no external stress on the cell. States denote the proteins being in particular domains, for example Tables 6.5 denotes the final state S1. The numbering of the states bears no relation to the the domains in which the states lie (not given). From the initial equilibrium state, the cell can transition through a number of regulatory (boxed) and switching (not boxed) states and lying in an attractor basis with the final equilibrium state in Table 6.5 as the attractor. Arrow directions denote allowed state transitions.

6.2.4 Results

Each set of initial conditions run through the model produces a set of states that the cell moves through. It was found that in each case there is a single attractor state that the system settles into, representing the decision of the cell to either continue normally, halt the cell cycle, signal for angiogenesis, or undergo apoptosis. An example of the transition diagram is shown in Figure 6.4. The set of four states, corresponding to the final states for each input 1-4, and the normal equilibrium positions in Table 6.2 are shown in Tables 6.5 through 6.8. For the alternate set of equilibrium positions in Table 6.3, the results are the same, except that when a single input is present, the

Table 6.5. Final state for no external stress on the cell. Final state for input set one: $(x_d, x_u) \in [0, \theta_d) \times [0, \theta_u)$. This represents no external stress on the cell. As expected, none of its responses are triggered, represented by $x_a < c_A$, $x_C < c_C$ and $x_D < c_D$.

Variable	Final domain
x_A	$[0, c_A)$
x_C	$[0, c_C)$
x_D	$[0, c_D)$
x_m	$[0, \theta_m)$
x_p	$[0, \theta_p)$

Table 6.6. Final state for UV stress on the cell. Final state for input set two: $0 \leq x_d < \theta_d$ and $\theta_u < x_u \leq m_u$. This represents one external stress on the cell, in this case UV stress. Two of its responses are triggered, represented by $c_A < x_a \leq m_A$, $c_C < x_C \leq m_C$ and $c_D < x_D \leq m_D$. In words, this means the cell is signalling for angiogenesis, and has halted the cell cycle in order to prevent DNA damage, but the cell is not under enough pressure to undergo apoptosis.

Variable	Final domain
x_A	$(c_A, m_A]$
x_C	$(c_C, m_C]$
x_D	$[0, c_D)$
x_m	θ_m
x_p	(θ_A, θ_D)

system only signals for the cell cycle to be halted and not angiogensis, and making $c_A < c_C$ reverses that. Also, for both inputs being present, the output is that both the cell cycle and angiogenesis are signaled for, but not apoptosis. This allows potentially harmful mutations to be passed on to daughter cells.

Table 6.7. Final state for DNA damage. Final state for input set three: $\theta_d < x_d \leq m_d$ and $0 \leq x_u < \theta_u$. This represents one external stress on the cell, in this case DNA damage. Two of its responses are triggered, represented by $c_A < x_a \leq m_A$, $c_C < x_C \leq m_C$ and $c_D < x_D \leq m_D$. In words, this means the cell is signalling for angiogenesis, and has halted the cell cycle to repair DNA damage, but the cell is not under enough pressure to undergo apoptosis.

Variable	Final domain
x_A	$(c_A, m_A]$
x_C	$(c_C, m_C]$
x_D	$[0, c_D)$
x_m	θ_m
x_p	(θ_A, θ_D)

Table 6.8. Final state for DNA damage and UV stress. Final state for input set four: $\theta_d < x_d \leq m_d$ and $\theta_u < x_u \leq m_u$. This represents two external stresses on the cell, DNA damage and UV stress. Two of its responses are triggered, represented by $c_A < x_a \leq m_A$, $c_C < x_C \leq m_C$ and $c_D < x_D \leq m_D$. In words, this means the cell is so stressed and damaged that it would be too risky for cell division to occur, which would fix in the genetic changes to its descendants, so it undergoes apoptosis.

Variable	Final domain
x_A	$(c_A, m_A]$
x_C	$(c_C, m_C]$
x_D	$[c_D, m_D)$
x_m	θ_m
x_p	$(\theta_A \theta_D)$

6.3 Mutations in *p53*

6.3.1 Methods

A binary string of given length, G, was used to represent each individual; each bit could be interpreted as being an allele or a base pair etc. A "0" was defined as being a healthy bit/allele and a "1" a mutated bit/allele. The phenotype of each genome was expressed as the number of mutations it contained (ie. the sum of "1" bits):

$$I_{\boldsymbol{x}} = \sum_{i=1}^{G} x_i, \tag{6.12}$$

where x_i is the i^{th} bit of the string $\boldsymbol{x}$.

The fitness function for each individual was simply the sum of "0" bits and could be expressed in terms of I_x, as:

$$F(\boldsymbol{x}) = G - I_{\boldsymbol{x}}. \tag{6.13}$$

The initial progeny were randomly created, where each bit of the array was set with a 90% probability of being healthy and 10% probability of being mutated.

A portion, g, of each genome was used to represent the *p53* gene. The size of g was set with respect to G. As discussed previously, the *p53* gene has numerous roles in maintaining a healthy cell cycle, such as cell cycle arrest, apoptosis and DNA repair (Vogelstein *et al.* 2000). In the model, the region representing the *p53* gene was designed such that 60 percent coded for apoptosis, g_1, and the other 40 percent for DNA repair, g_2.

The evolution of the progeny was studied over fifty cell cycles. During each cell cycle, all individuals underwent conditional spontaneous mutation and DNA repair (if still functioning). Cells underwent apoptosis if they were sufficiently mutated but the apoptosis part of the *p53* gene was still functioning. All of this is described in more detail below.

The model considered both silent and missense base pair substitution mutations. The virtual cells were mutated with mutation rate m per bit per cell cycle. The rate of spontaneous mutations for *E. coli* was stated by Elliott to be one error in every 10^8 nucleotides replicated (Elliott and Elliott 1997). Other rates of mutation have been found to be in the order of one nucleotide change per 10^9 nucleotides per cell generation for other bacteria (Alberts *et al.* 2002) and somewhere in the order of one in 10^6 or 10^5 nucleotides replicated for humans (Elliott and Elliott 1997). Since the model considered only genomes of one hundred bits and the number of generations it considered was limited due to the time required to perform the computations, using mutation rates of $1/10^6$ over a limited number of generations would have produced little or no effect on the population. To compensate for this, much larger mutation rates, in the range of 0.05 to 0.50, were used.

For individuals in the model to be eligible for DNA repair, a minimum fitness was required of the g_2 region. Each individual that satisfied this minimum fitness level underwent DNA repair, with a repair rate m_1. Again, this rate had no reflection on the actual rate of repair occurring in cells. The range of values used for m_1 was the same as those used for m. Various relationships between m and m_1 were used, such as $m_1 = m$, $m_1 = 2m$, $m_1 = 10m$, etc. When cells underwent DNA mutation or repair, each bit/allele in its genome had an equal likelihood of experiencing a mutation.

If an individual did not meet the minimum fitness requirement then apoptosis of the cell was required. For apoptosis to occur, however, a minimum fitness of the g_1 region was also essential. Individuals, which required apoptosis because $F(x) < f$, and that had a g_1 region with minimum fitness, $F(g_1) > p_1$, were then selected for apoptosis. If the g_1 region was too mutated, then apoptosis would not occur and the severely mutated cell would survive and could then undergo cell division to produce two daughter cells with extremely mutated genomes.

Cell division occurred at the end of each cell cycle, so that each successive generation had twice the population size of its parent generation. To ensure the cells did not proliferate out of control, random individuals were killed off directly after cell division so that the population size was kept to a maximum of ten thousand.

6.3.2 Results

The genome size was set to $G = 100$, with $g = 8$ and the initial population size set to $N = 50$. For cells to survive, it is essential that the rate of repair exceeds the rate of mutation, so initially the mutation rate was kept constant at $m = 0.05$ and the following repair rates were used: $m_1 = m = 0.05$; $m_1 = 2m = 0.1$; $m_1 = 4m = 0.2$ and $m_1 = 10m = 0.5$. Due to the random element within our model, ten trials were run for each of these repair rates, recording the mean phenotype (number of mutations) of the progeny at each generation.

The mean phenotype $\bar{I}_x(t)$ was then plotted against time t (in cell cycles), for each repair rate considered, and the resultant plot in Figure 6.5 shows a second order polynomial relationship between the mean phenotype $\bar{I}_x(t)$ and time t.

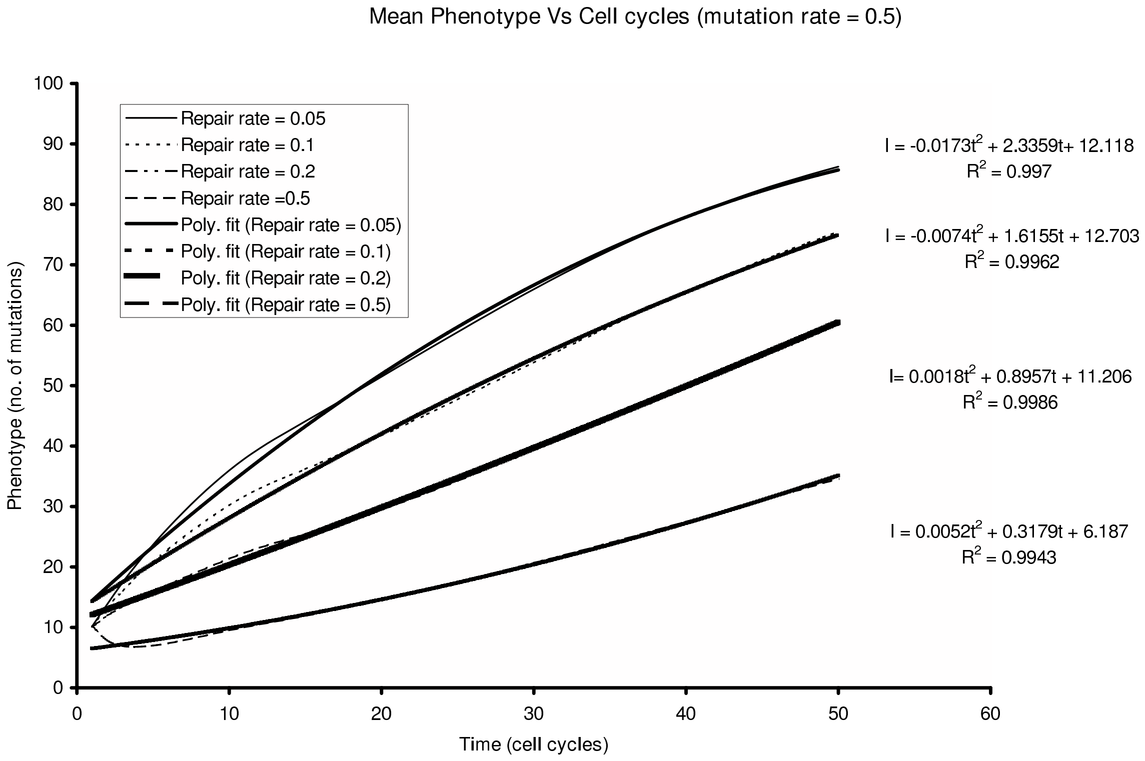


Figure 6.5. Mean phenotype (number of mutations) of cells against generations. The mutation rate $m = 0.05$. Repair rates considered: $m = 0.05$ $m = 0.1$ $m = 0.2$ and $m = 0.5$. The actual curves are represented by the four different line types. A cubic polynomial was fitted to each of the curves and these are shown in bold. The equations for each of the fitted polynomials appears to the right of the respective curve, along with the corresponding R^2.

From these plots, one can see that the phenotype of the population continues to increase with time, for all the repair rates. In other words, if one continued to observe the genomic activity of the cells, all the cells in the progeny would eventually acquire a phenotype of 100 (i.e. 100 mutations); the difference between the curves, however, is a function of the rates of mutation acquisition. The latency to reach this level of mutation would be different for the different repair rates.

Table 6.9. Mean phenotype and its derivative for different repair rates. This table shows the mean phenotype $\bar{I}_x(t)$ and its derivative $\bar{I}'_x(t)$ for various repair rates (mutation rate = 0.05).

Repair rate	$\bar{I}_x(t)$	corresponding R^2	$\bar{I}_x(t)'$
0.05	$-0.0173t^2 + 2.3359t + 12.118$	$R^2 = 0.997$	$-0.346t + 2.3359$
0.1	$-0.0074t^2 + 1.6155t + 12.703$	$R^2 = 0.9962$	$-0.148t + 1.6155$
0.2	$0.0018t^2 + 0.8957t + 11.206$	$R^2 = 0.9986$	$0.0036t + 0.8957$
0.5	$0.0052t^2 + 0.3179t + 6.187$	$R^2 = 0.9943$	$0.0104t + 0.3179$

Differentiating $\bar{I}_x(t)$ with respect to t one can see the rate of mutation acquisition per cell division. Table 6.9 shows the mean phenotype $\bar{I}_x(t)$ and its derivative $\bar{I}'_x(t)$ for each curve in Figure 6.5.

Plotting the derivative $\bar{I}'_x(t)$ for each mutation rate in Figure 6.6, we see that the curves for $m_1 = 0.05$ and $m_1 = 0.1$ both have a decreasing mutation acquisition rate, but the curves for $m_1 = 0.2$ and $m_1 = 0.5$ both have an increasing rate of mutation acquisition.

From these results, one can hypothesize that the rate of mutation acquisition is of the form: $ax^2 + bx + c$ (and concave therefore $a < 0$). Integrating this back with respect to t, to give the expression for the mean phenotype, would infer that the curves representing the behaviour of the phenotype over time are in fact of the form: $ax^3 + bx^2 + cx + d$, rather than $ax^2 + bx + c$. The rate of mutation acquisition is at first slow and increasing, up to a threshold point; after this point, the cell continues to gain mutations, but the rate of acquisition decreases. If this theory is correct, then it would follow that the first two curves are in fact past the threshold point at time $t = 0$. The turning point at which the mutation acquisition rate goes from increasing to decreasing occurs at some point $t < 0$ and hence does not appear. Since the second two curves ($m_1 = 0.2$ and $m_1 = 0.5$) still have an increasing rate of mutation acquisition, it would suggest that this "threshold" point has not yet been reached. If one was to continue to observe the behavior of the two curves $m_1 = 0.2$ and $m_1 = 0.5$, they would reach their "threshold point" and the rate of mutation acquisition would decrease until the entire genome was mutated (i.e. $\bar{I}_x(t) = 100$).

To test this theory, the case $m = 0.05$, $m_1 = 0.5$ was considered, and the evolution of the cell population through 200 cell cycles was followed. The resultant plot is shown

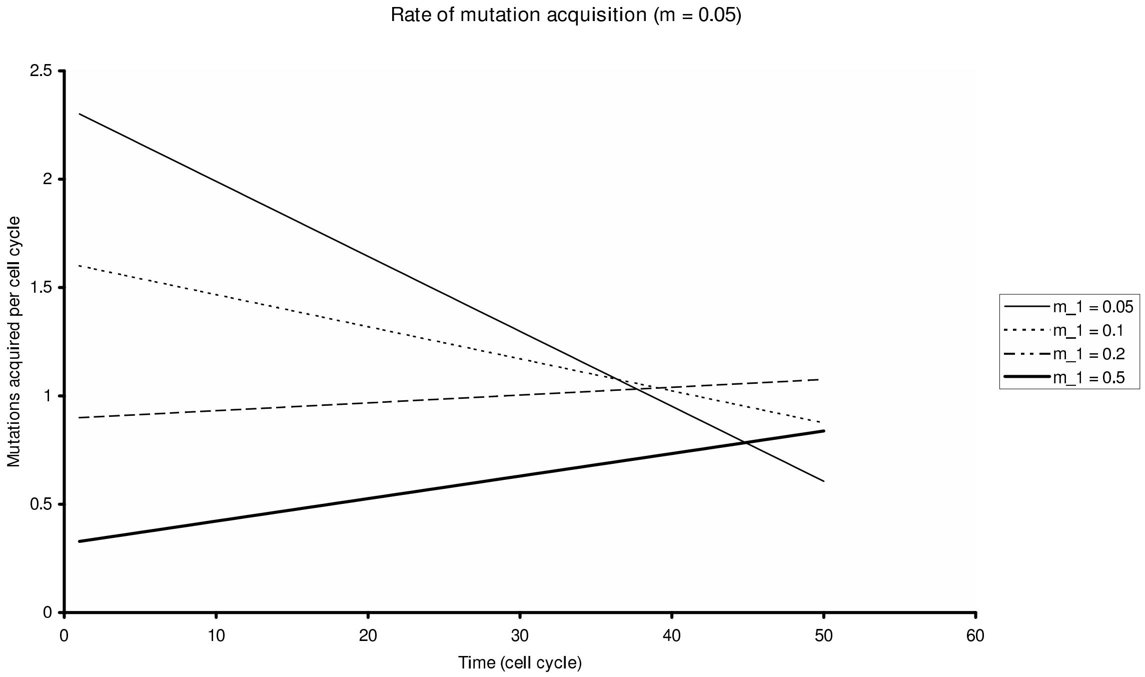


Figure 6.6. Rate of mutation acquisition per cell cycle. The mutation rate $m = 0.05$. Repair rates considered are $m_1 = 0.05$, $m_1 = 2m = 0.1$, $m_1 = 4m = 0.2$, and $m_1 = 10m = 0.5$. The equations from Figure 6.5, for the fitted polynomials, were differentiated with respect to t and plotted. Each line shown displays the mean rate of mutation acquisition at time t for the respective repair rates considered.

in Figure 6.7. As can be seen from the plot, the function for this curve is a cubic polynomial (with $R^2 = 0.9984$), which supports the theory that the rate of mutation would continue increasing to a point and then begin to decrease. Differentiating this and solving for $\bar{I}_x(t) = 0$ will give the cell cycle at which the turning point occurs. Note that this doesn't make sense on an individual cell basis, but does when one is considering the population mean. A different function fitted (such as the logistic function below) may not have such a turning point, however the cubic polynomial curve seems a good approximation.

This procedure was repeated, this time keeping the repair rate constant at $m_1 = 0.05$ and considering mutation rates $m = 0.05$, $m = 0.01$ and $m = 0.5$. The mean phenotype was then plotted against time t (in cell cycles), for the three different mutation rates considered (Figure 6.8). Once again one can see large rates of mutation acquisition at the beginning, which decrease with time. Once the entire population has a fully mutated genome the individuals in the population can no longer acquire mutations and consequently the mutation acquisition rate will be zero.

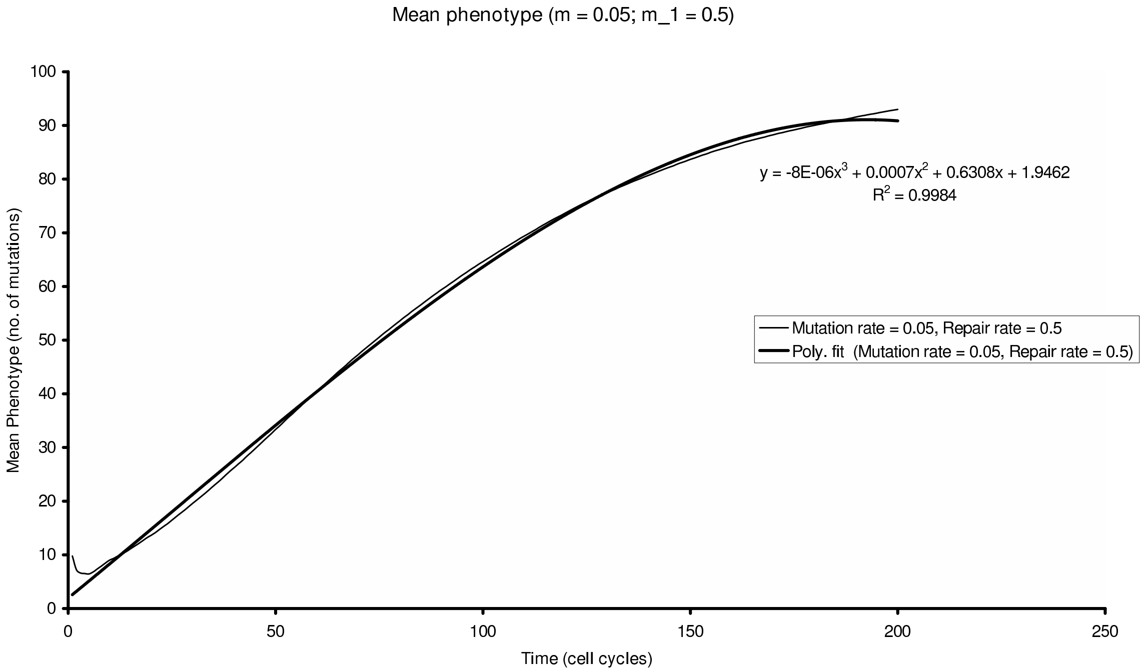


Figure 6.7. Plot of genomic activity. The genomic activity of the case with mutation rate $m = 0.05$ and repair rate $m_1 = 0.5$ was observed over 200 cell cycles. The actual curve is the fine line and the bold line over the top is the fitted polynomial. The equation for the polynomial is shown, with its R^2 value. Watching the evolution over a longer time period showed that the shape of the curve changes from convex to concave. The point of inflection is the threshold point where the rate of mutation acquisition changes from increasing to decreasing.

The curves representing the mean phenotype of the population may be better expressed using the logistic differential equation (Boccara 2003). This model is based on the assumption that a population grows at a rate proportional to the size of the population and is expressed as:

$$\frac{d\bar{I}}{dt} = k\bar{I}\left(1 - \frac{\bar{I}}{K}\right), \tag{6.14}$$

where $\bar{I}$ is the mean phenotype/number of mutations, k is the proportionality constant, and K is the carrying capacity.

For the model, the carrying capacity K would be 100—the length of the genome, G—and the proportionality constant k would be determined by the ratio between the mutation and repair rates, (m and m_1 respectively).

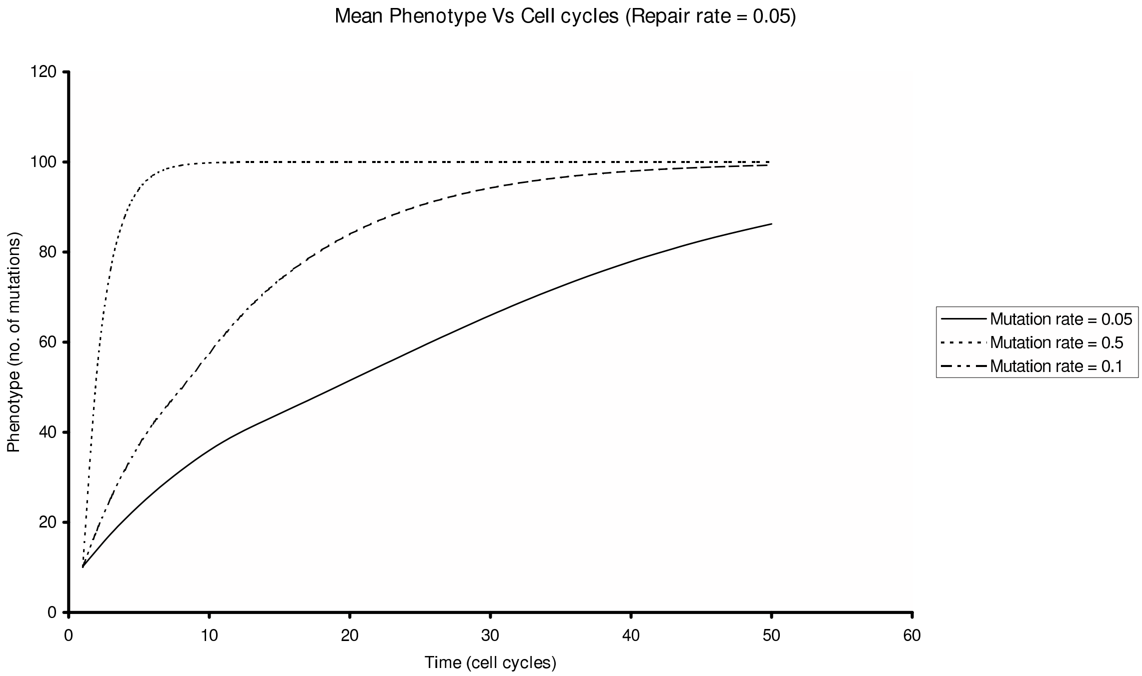


Figure 6.8. Mean phenotype (number of mutations) of cells against generations (cell cycles). Repair rate $m = 0.05$. Mutation rates considered are $m = 0.05$, $m = 0.1$, and $m = 0.5$. The concave shape of the curves indicates that the rate of mutation acquisition is decreasing, and hence the inflection points would have occurred at time $t < 0$. For all curves, mutations are acquired at a decreasing rate until a maximum phenotype of 100 is obtained.

6.3.3 Discussion

The results obtained suggest that the mean phenotype $\bar{I}_x(t)$ with respect to time is determined by a cubic polynomial of t. The rate of mutation acquisition increases to a point, before decreasing as the mean phenotype of the progeny approaches a maximum.

In healthy cells, the acquisition of mutations is kept in check by the repair mechanisms of the cell. This is best represented by the curve in Figure 6.5 where the rate of repair is ten times the rate of mutation ($m_1 = 0.5$, $m = 0.05$). There is a greater latency in obtaining mutations, when compared to all the other cases (i.e. a significantly lower rate of mutation acquisition). Over time, however, more and more mutations are acquired and the repair mechanisms of the cell become less efficient, allowing the rate at which these mutations are acquired to increase, up to a point. Past this point, as the cell becomes more and more mutated, it is no longer able to sustain this increasing rate of mutation. Consequently, this rate drops off, until the entire genome is mutated. This process varies stochastically, and this was modelled, resulting in a heterogeneous population, however since cell division and cell-cell interactions were not modelled, no conclusions can be drawn as to the spatial nature of tumour progression.

Cells in the human body behave in a similar manner. Over time, they acquire mutations, which are ordinarily kept in check by repair mechanisms, such as those initiated by the *p53* gene. As cells age, however, the efficiency of the repair mechanisms decreases and they develop more and more mutations. As the mutations in a cell accumulate, the cell can become malignant, and a tumour will result. In Figure 6.8, with the mutation rate ten times the repair rate, the repair rate of the cell has virtually no impact on keeping the mutation rate in check and the progeny of cells accumulate a mean phenotype of 100 after only 10 cell cycles.

Future research should consider using more realistic mutation and repair rates, and the interactions of gene and drug therapies on these, to determine their usefulness and efficacy. Another interesting question is: given the mutations in *p53*, what are the benefits and drawbacks in using radiotherapy, which while killing off some cells will further mutate others?

6.4 Conclusions

6.4.1 Gene networks

Given a qualitative model of the interactions of the p53 protein in the cell, a simplified gene network was designed to represent this information. In a qualitative sense this gives the output that is observed in *in vivo* and *in vitro* cells, namely that if not stressed or damaged, the cell undergoes a normal cell cycle and divides. If there is some damage or stress the cell cycle is halted, and for a significant amount of damage and stress the cell does not divide and pass on its mutations but instead undergoes apoptosis (programmed cell death). Shifting the equilibrium positions is equivalent to altering the response of the cell; for example, moving the equilibrium positions lower means the cell is less likely to signal for appropriate responses, in particular never undergoing apoptosis. This would mean that potentially harmful mutations, from the perspective of the organism, are carried forward into daughter cells upon division of the stressed and damaged cell. Shifting the equilibrium positions higher means the cell is more likely to halt and repair DNA and undergo apoptosis for higher (yet lower than normal) levels of DNA damage. This could be a useful response to have in tumors that have not lost the capability of apoptosis and suggests future treatments could target these equilibrium positions through drugs altering the chemical balance and thus the threshold values in signalling networks.

In future work on the *p53* gene network, the angiogenesis, cell cycle, and apoptosis pathways could be studied in more detail. The use of quantitative data where available would allow one to build testable models of cell outputs, although the existing model presented herein agrees with observed qualitative descriptions of cell processes. In addition to exploring more equilibrium positions and inputs, changes to the network could be explored. Since many drugs alter the pathways, this research could possibly impact upon drug selection, dosage, and development of new drugs for treating cancer.

6.4.2 Modeling *p53* mutations

The model of *p53* mutations offers a simplistic view of the effects of mutation on a population of cells. The model considers the role of *p53* in maintaining genomic integrity, and investigates how mutations to this gene affect the evolution of a population of cells. It is observed that the repair mechanisms of the cell are not able to prevent mutations of the cell indefinitely, but instead increase the latency of mutation acquisition. Because the efficiency of the repair mechanism in cells decreases with age, all replicating cells are predisposed to potentially developing into a tumour with further mutations fixed in upon replication (Spencer *et al.* 2004). Methods that can prolong the time that mutation of the cell is kept in check will reduce the potential of tumour development. The results show that increasing the ratio between the repair and mutation rates increases the ability of the cell to keep mutations in check for longer. Any future gene therapies or drug treatments could potentially alter these rates in order to prolong life or possibly even prevent tumours progressing through to cancer.

6.4.3 General conclusions

The *p53* gene plays an important role in preventing the accumulation of genetic damage. Loss of its functionality can have a drastic impact on the mutation rate in cells, and this is explored in the following chapter on cancer. The cell fate (apoptosis or repair) that p53 selects for is dependent on the types of damage. It was repaired if only one of UV or general (for example, chemical) damage occurs, but if multiple causes of damage occur then apoptosis is signalled for. There could be more research done on combining these two models to look at mutations in the presence of various carcinogens.

To summarise the novel contributions of this work: a switching network model of the p53 gene network was developed, and the output of this network was explored, given such inputs as UV-induced and general DNA damage. Mutations in p53 and their effect on DNA repair processes, and hence on the time to onset of cancer, were explored.

Chapter 7

Cancer

CANCER is viewed as a multistep process whereby a normal cell lineage is transformed into a cancer cell lineage through the acquisition of mutations. In this chapter, the complexities of cancer progression are reduced to a simple set of underlying rules that govern the transformation of normal cells through to malignant cells. In doing so, an ordinary differential equation model is derived that explores how the balance of angiogenesis, cell death rates, genetic instability, and replication rates give rise to different kinetics in the development of cancer. The key predictions of the model are that cancer develops fastest through a particular ordering of mutations and that mutations in genes that maintain genomic integrity would be the most deleterious type of mutations to inherit. In addition, a sensitivity analysis is performed on the parameters included in the model to determine the probable contribution of each. This chapter presents a novel approach to viewing the genetic basis of cancer from a systems biology perspective and provides the groundwork for other models that can be directly tied to clinical and molecular data.

7.1 Introduction

The standard perspective on cancer progression is that it is a form of somatic evolution where certain mutations give one cell a selective growth advantage (Cahill *et al.* 1999). Oncogenesis[2] is thought to require several independent, rare mutation events to occur in the lineage of one cell (Nowell 1976). Kinetic analyses have shown that four to six rate-limiting stochastic mutational events are required for the formation of a tumour (Armitage and Doll 1954, Renan 1993). Hanahan and Weinberg (2000) proposed the following six hallmark capabilities that normal cells must acquire to become a cancerous: (i) self-sufficiency in growth signals, (ii) insensitivity to anti-growth signals, (iii) evasion of apoptosis, (iv) limitless replicative potential, (v) sustained angiogenesis, and (vi) tissue invasion and metastasis. They define genetic instability as an "enabling characteristic" that facilitates the acquisition of other mutations due to defects in DNA repair processes. These characteristics are simplified for the purposes of modelling to the following four: angiogenesis (A), immortality, including evasion of cell death (D), genetic instability, a function of mutation rates (G), and increased replication rate (R). Invasion and metastasis (M) is considered as a final step that allows the spread of a localized tumour. In line with the views of Hanahan and Weinberg (2000), cancer research is developing into a logical science where the molecular and clinical complexities of the disease will be understood in terms of a few underlying principles. The multistep progression to cancer is explored using an ordinary differential equation (ODE) model, which, despite the apparent complexity of the equations, is based on basic principles and a minimal set of parameters.

7.1.1 Novel contributions

The novel contributions of this work, carried out in collaboration with Sabrina L. Spencer and José A. García are:

1. Development of a mathematical model describing, in a general sense, the progression of normal to cancerous tissue.
2. Exploration of various parameters (such as mutation rates) on the average time to development of cancer.

[2] For definitions of key terms, refer to the thesis conventions and glossary in the front matter and also Appendix A.

3. Exploration of the kinetics of various genetic pathways to cancer, for example, do more cells encounter genetic instability before loss of programmed cell death functionality, or after?

4. The affect of inheriting mutations on the mean age at which cancer occurs.

My key contributions to this work were in:

- Helping develop the equations.
- Development of the general form of the equations and numerical software solutions to this.
- Assisting with the interpretation of results.
- Deriving formulae for establishing the range of cell birth and death rates to be explored in the sensitivity analysis.

7.2 Structure and parameters of the model

Although the model applies to the process of oncogenesis in general, the parameters are loosely based on breast cancer data. The following cell populations are considered: a population of 10^8 normal cells (N), cells which have acquired the ability to induce angiogenesis (A), cells with mutations which allow them to avoid death (D), cells with mutations that lead to genetic instability (G), cells with mutations which increase their replication rate (R), and cells with two or more of these mutations. Cell populations that have acquired two or three mutations are denoted by listing the mutations together in alphabetical order (state DRA would be listed as state ADR, for example). A cell which has acquired all four mutations is labelled a primary tumour cell (T). Although the model only addresses the development in genetic detail, the migration of tumour cells to other locations in the body and subsequent site invasion is considered a rate-limited step in the model, with such metastatic cells labelled M.

The spontaneous mutation rate in human cells has been estimated to be in the range of 10^{-7} to 10^{-6} mutations /gene / cell division (Jackson and Loeb 1998). A spontaneous mutation rate of of $k_1 = 10^{-7}$ mutations / gene / cell division is assumed. The loss of DNA repair genes can increase the mutation rate by a factor ranging from 10^1

to 10^4 (Tomlinson *et al.* 1996). It is assumed that the mutation rate after a genetic instability mutation increases 1000-fold to $k_2 = 10^{-4}$ mutations / gene / cell division. Successful invasion and metastasis depend upon acquisition of the other hallmark capabilities, as well as several new capabilities (Hanahan and Weinberg 2000). To simplify the model, the multistep progression of a tumour cell to a metastatic cell, this complex process is considered as a single step. It has been estimated that the rate of successful metastasis is in the range of 10^{-9} to 10^{-7} per cell division (Luebeck and Moolgavkar 2002), and so a conservative estimate of $k_3 = 10^{-9}$ was used for the rate of transition from a primary tumour cell to a metastatic cell.

A tumour cannot grow past about 10^6 cells without angiogenesis supplying blood to the tumour (Folkman 1990). Thus the size of the tumour is capped at 10^6 cells until greater than 10% of the population of non-normal, non-metastatic cells have acquired a mutation in an A gene. This accounts for the fact that only a fraction of the cells in a tumour need to send angiogenesis signals in order to develop an adequate blood supply for the tumour. In addition, populations of non-normal, non-metastatic cells are always capped by a lethal tumour burden limit of 10^{13} cells (Friberg and Mattson 1997), irrespective of the angiogenesis cap.

Futreal *et al.* (2004) state that 291 genes have been reported to be implicated in the causation of human cancer and note that many more cancer genes remain to be identified. Thus an estimate of 400 genes involved in the development of a primary tumour was used. The number has It is assumed there are 100 genes involved transitions where only one mutation is acquired (e.g. $N \rightarrow A$, $D \rightarrow DR$, $AG \rightarrow AGR$, or $ADG \rightarrow ADGR$), 10 genes involved in transitions where two mutations are acquired in one step (e.g. $N \rightarrow AD$, $G \rightarrow ADG$, or $AR \rightarrow ADGR$), and 1 gene involved in transitions where three mutations are acquired in one step (e.g. $N \rightarrow ADG$ or $G \rightarrow ADGR$). This feature accounts for a mutational hit in *p53*, for example, which could take a cell directly from N to DGR, as p53 is involved in apoptosis, DNA repair, and cell cycle progression (Vogelstein *et al.* 2000). More information on the cell cycle can be found in Appendix A and the biology of mutation was discussed in Chapter Six.

An estimate that the relative contribution to increased net proliferation for mutations in the D and R categories is 7 and 3, respectively, an inference made from work by Tomlinson and Bodmer (1995). Using this D:R ratio of 7:3, a tumour volume doubling time for breast cancer of 500 days (Friberg and Mattson 1997), and a cell division rate

Table 7.1. Cancer model parameters. Parameters appearing in the ODE model. The default value is that used in the ODEs, unless otherwise specified.

Characteristic	Parameter	Range in literature	Default value	Reference
Mutation rate without a G mutation	k_1	10^{-7}–10^{-6} mut./gene/cell div.	10^{-7}	Jackson and Loeb (1998)
Mutation rate with a G mutation	k_2	10^{-6}–10^{-2} mut./gene/cell div.	10^{-4}	Tomlinson *et al.* (1996)
Metastasis rate	k_3	10^{-9}–10^{-7} /cell division	10^{-9}	Luebeck and Moolgavkar (2002)
Number of genes involved in cancer		291+ genes	400	Futreal *et al.* (2004)
Genes per single, double, triple transitions		unknown	100, 10, 1	N/A
Tumour volume doubling time		88–523 (sometimes ¿5000) days	500	Friberg and Mattson (1997)
Relative contribution of D:R		7:3–8:2 (inferred)	7:3	Tomlinson and Bodmer (1995)
Cell division rate without an R mutation	$1/b$	once every 1.8-47.5 days	$1/10.00$ days^{-1}	Rew and Wilson (2000)
Cell division rate with an R mutation	$1/b_R$		$1/9.92$ days^{-1}	see section 7.4
Cell death rate without a D mutation	$1/d$	once every 1.8-47.5 days	$1/10.00$ days^{-1}	Rew and Wilson (2000)
Cell death rate with a D mutation	$1/d_D$		$1/10.11$ days^{-1}	see section 7.4
% of cells needed to signal for A		unknown	10%	N/A
Angiogenesis cap		10^6 cells	10^6	Folkman (1990)
Lethal tumour burden cap		10^{13} cells	10^{13}	Friberg and Mattson (1997)

for breast cancer of $1/10.00$ days^{-1} (Rew and Wilson 2000), the following are calculated (see Section 7.4 for formulae): cells without a mutation in an R gene divide every $b = 10.00$ days, cells with a mutation in an R gene divide every $b_R = 9.92$ days, the lifetime of cells without a mutation in a D gene is $d = 10.00$ days, and the lifetime of cells with a mutation in a D gene is $d_D = 10.11$ days. The birth and death rates are equal for normal cells and for all cells without a mutation in D or R.

The above information is depicted in a unified fashion in Figure 7.1 and the parameters appearing in the ODE model are given in Table 7.1.

In the next section, an ODE model is presented that was used to explore the following areas:

1. The kinetics of various paths to cancer.
2. The effect of inherited mutations on cancer development.
3. A sensitivity analysis of variations in the parameters.

7.3 Construction of the ODE model

Although the development of cancer has inherent stochasticity and several previous cancer models include stochastic components (Speer *et al.* 1984, Koscielny *et al.* 1985), ordinary differential equations can model the mean of the processes associated with cancer progression. The goal was to create one "generic" model to better understand

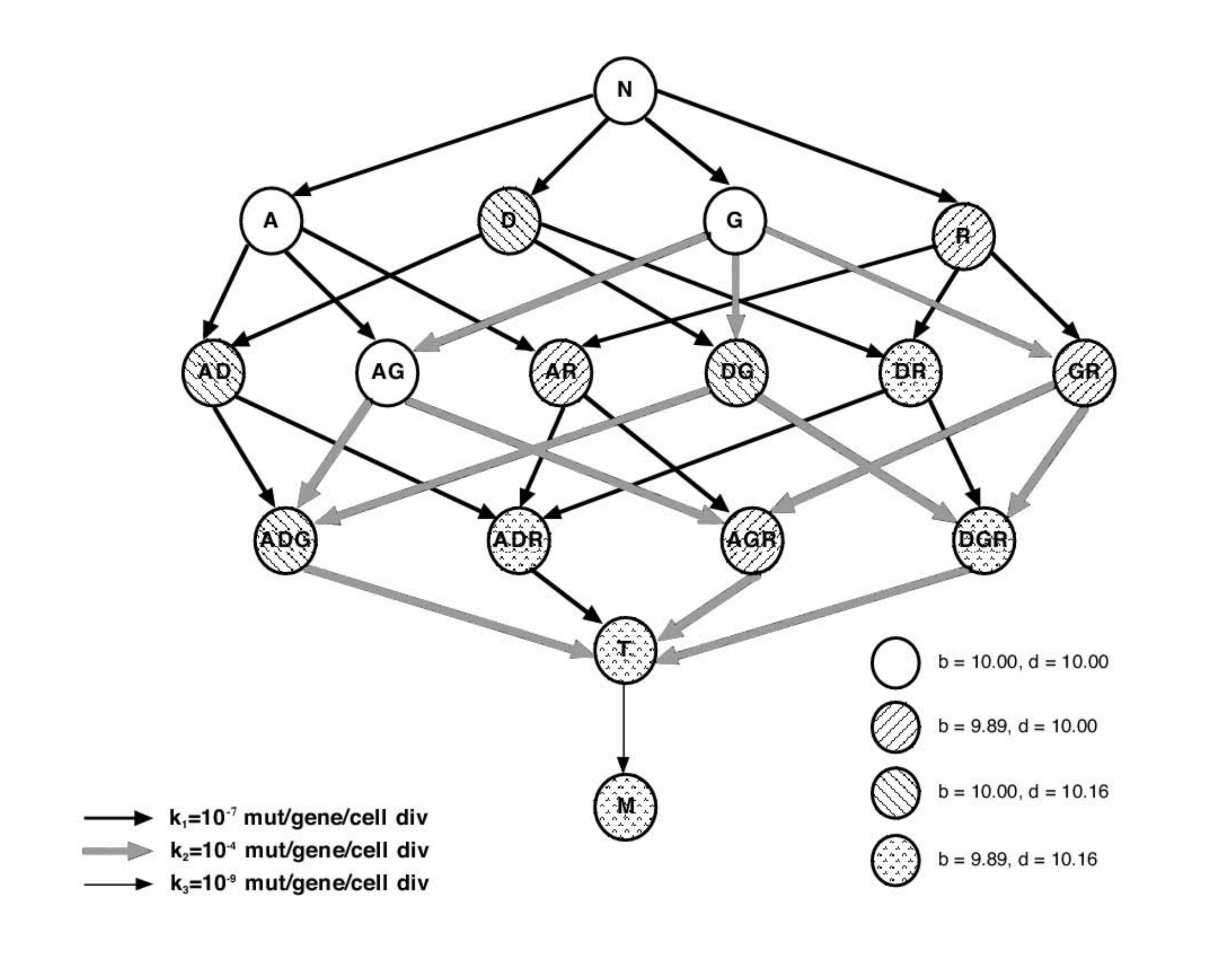


Figure 7.1. State diagram of the cancer model. Normal cells (N) can acquire mutations which give the cell the capability to induce angiogenesis (A), mutations which give the cell the capability to avoid death (D), mutations which lead to genetic instability (G), or mutations which increase the proliferation rate (R). These mutations are acquired at rate k_1. After a mutation in G, the mutation rate increases to k_2. Cells with one mutation go on to acquire two, three, and four mutations, denoted by listing the mutations together in alphabetical order for the cases of two and three mutations. When a cell has acquired all four mutations, it becomes a primary tumour cell (T). Finally, tumour cells become metastatic cells (M) at rate k_3. Double and triple state transitions are also allowed, as detailed in the text, but are not shown in this diagram for simplification. Cell birth rates $(1/b)$ and cell death rates $(1/d)$ have units days^{-1}.

the kinetics of cancer progression, not to create a model that captures the variability of many different types of cancer all at once.

Based on the basic rules outlined in the state diagram in Figure 7.1, 17 ODEs are constructed to model a heterogeneous population of cells undergoing the multistep process of tumourigenesis. Each equation represents one of the 17 populations of cells depicted in the state diagram and has the following format: the population of cells in a state is increased by cells gaining mutations and entering that state from previous states, is increased by cells replicating and remaining in that state, and is decreased by cells gaining new mutations and leaving that state for a new state. The populations are capped by two logistic terms, as detailed below.

The ODEs can be condensed into vector format as follows:

$$\begin{aligned}\frac{d\boldsymbol{y}}{dt} = \left(\operatorname{diag}\left(\operatorname{diag}\left(\boldsymbol{y}^T\boldsymbol{k}\right)^T \boldsymbol{b}\right)\boldsymbol{M} + \operatorname{diag}\left((\boldsymbol{b}-\boldsymbol{d})^T\boldsymbol{y}\right)\right)\boldsymbol{S}\left(1 - a(\boldsymbol{y})\frac{P_{\overline{NM}}}{10^6}\right)& \\ \times\left(1-\frac{P_{\overline{NM}}}{10^{13}}\right) + \boldsymbol{m_m},&\end{aligned} \tag{7.1}$$

where $\boldsymbol{y}$ is the row vector of cell populations; y_1 is the population of normal cells, $y_2, y_3, \ldots, y_{15}$ are the populations of cells with single, double, and triple mutations, y_{16} is the number of primary tumour cells (cells with all four mutations), and y_{17} is the number of metastatic cells. Here, $\operatorname{diag}(\cdot)$ is the operator which forms the row vector of the main diagonal of the matrix. The corresponding rate (row) vector is $\boldsymbol{k}$, with mutation rates k_i (mutations / gene / cell division) corresponding to the mutation rate for element y_i in $\boldsymbol{y}$. The same applies to the birth rates $\boldsymbol{b}$ (day^{-1}) and death rates $\boldsymbol{d}$ (day^{-1}). The metastasis rate vector is $\boldsymbol{m_m} = \left(0, 0, \ldots, 0, 10^{-9} \times y_{16}\right) + (1/b_R - 1/d_D)y_{17}$, corresponding to cells leaving y_{16} for y_{17} at rate 10^{-9}, and a doubling of metastatic cells at rate $(1/b_R - 1/d_D)$ for $1/b_R$ and $1/d_D$ as given in Table 7.1. The 17×17 upper triangular matrix $\boldsymbol{M}$ consists of elements $M_{i,j}$ $(j \neq i)$ for the number of genes associated with going from state i to state j, and

$$M_{i,i} = -\sum_{j \neq i} M_{i,j}, \tag{7.2}$$

is the main diagonal containing the number of genes for leaving each of the states. $\boldsymbol{S}$ is the 17×17 matrix

$$\boldsymbol{S} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & & 1 & 0 \\ 0 & \dots & & 0 & 0 \end{pmatrix} \tag{7.3}$$

used to apply the cell population caps to the non-normal, non-metastatic cells. Non-normal, non-metastatic cells are denoted by $P_{\overline{NM}}$, where

$$P_{\overline{NM}} = \left(\sum_{i=2}^{16} y_i \right). \tag{7.4}$$

The system is capped at 10^6 cells using a logistic term if 10% or fewer of the non-normal, non-metastatic cells are in states with angiogenesis mutations, otherwise this term is removed. This is expressed in the term $a\left(\boldsymbol{y}\right)$, defined as

$$a\left(\boldsymbol{y}\right) = \begin{cases} 0, \frac{P_A}{P_{\overline{NM}}} > 10\% \\ 1, \text{otherwise}, \end{cases} \tag{7.5}$$

where P_A is the number of non-metastatic cells with mutations in A category genes. The populations of non-normal, non-metastatic cells are also capped by a lethal tumour burden limit of 10^{13} cells (Friberg and Mattson 1997), irrespective of the angiogenesis cap. The ODEs are solved using the Runge-Kutta method of order 5 (Press *et al.* 1986), with a variable step size between 1 and 10^{-5}, to guarantee the errors in calculating the populations remain within 10^{-4}.

Below, the normal cell and four single state ODEs are unpacked from the compact vector form for ease of comprehension. Note that the equations assume a constant, renewing population of normal cells, since it is assumed that cells leaving state N for other states are few enough in number so as not to affect the population of N cells.

$$
\begin{aligned}
\frac{dP_N}{dt} &= 0, && \text{(7.6a)}\\
\frac{dP_A}{dt} &= \left(\frac{100P_N k_1}{b} - \frac{(3\times 100 + 3\times 10 + 1)\,P_A k_1}{b}\right)\left(1-\frac{P_{\overline{NM}}}{10^6}\right)\left(1-\frac{P_{\overline{NM}}}{10^{13}}\right), && \text{(7.6b)}\\
\frac{dP_D}{dt} &= \left(\frac{100P_N k_1}{b} + P_D\left(\frac{1}{b}-\frac{1}{d_D}\right) - \frac{cP_D k_1}{b}\right)\left(1-\frac{P_{\overline{NM}}}{10^6}\right)\left(1-\frac{P_{\overline{NM}}}{10^{13}}\right), && \text{(7.6c)}\\
\frac{dP_G}{dt} &= \left(\frac{100P_N k_1}{b} - \frac{cP_G k_2}{b}\right)\left(1-\frac{P_{\overline{NM}}}{10^6}\right)\left(1-\frac{P_{\overline{NM}}}{10^{13}}\right), && \text{(7.6d)}\\
\frac{dP_R}{dt} &= \left(\frac{100P_N k_1}{b} + P_R\left(\frac{1}{b_R}-\frac{1}{d}\right) - \frac{cP_R k_1}{b_R}\right)\left(1-\frac{P_{\overline{NM}}}{10^6}\right)\left(1-\frac{P_{\overline{NM}}}{10^{13}}\right), && \text{(7.6e)}\\
&\vdots
\end{aligned}
$$

Note that $c = (3\times 100 + 3\times 10 + 1)$ denotes the number of ways of getting into that state from mutations in any of the genes in any of the previous set of states.

The equations for the other populations follow the same format and can be derived from the state diagram and from the vector form of the ODEs. In words, Equation 7.6c, for example, says that the population of cells with a mutation in D is increased by normal cells gaining a mutation in one of 100 genes in D at a rate of k_1 every b days. The population is also increased by cells in state D replicating (but not mutating) every b days and dying every d_D days. The population is decreased by cells leaving state D and gaining a single mutation in one of 3 other categories (AD, DG, DR) each with 100 genes, by gaining a double mutation in one of 3 ways (DGR, ADG, ADR), with 10 genes being involved in each transition, or by gaining a triple mutation to go to state $ADGR$ with 1 gene being involved in the transition. The logistic term caps the total population of non-normal, non-metastatic cells at 10^6 cells. What is not visible in this standard form of the ODEs but is present in the vector form is the fact that the logistic angiogenesis cap is only imposed when 10% or fewer of the non-normal, non-metastatic cells have a mutation in the A category. Finally, populations of non-normal, non-metastatic cells are always capped by a lethal tumour burden limit of 10^{13} cells.

7.4 Calculation of cell division and cell death rates

In order to calculate the change in cell division and cell death rates for mutations in R and D, start with the assumption that birth and death rates are equal for normal cells, that is, the cell division rate = $1/b$ = cell death rate = $1/d = 1/10$ days^{-1}.

The ODEs are then approximated by equations where the rate of cells entering and leaving the state are considered as negligible compared with the tumour volume doubling time, since they are several orders of magnitude different. Thus, for cells with mutations in D but not R, say, consider

$$\frac{dD}{dt} = D\left(\frac{1}{b} - \frac{1}{d_D}\right). \tag{7.7}$$

Solving this gives $D = D_0 \exp\left((1/b - 1/d_D)\, t\right)$. A doubling corresponds to $2 = \exp\left((1/b - 1/d_D)\, T_D\right)$. Similarly, for cells with mutations in R and not D, one arrives at $R = R_0 \exp\left((1/b_R - 1/d)\, t\right)$. Taking the natural logarithm of both sides leads to Eq's 7.8a and 7.8b,

$$\frac{\ln 2}{1/b_R - 1/d} = T_R, \tag{7.8a}$$

$$\frac{\ln 2}{1/b - 1/d_D} = T_D, \tag{7.8b}$$

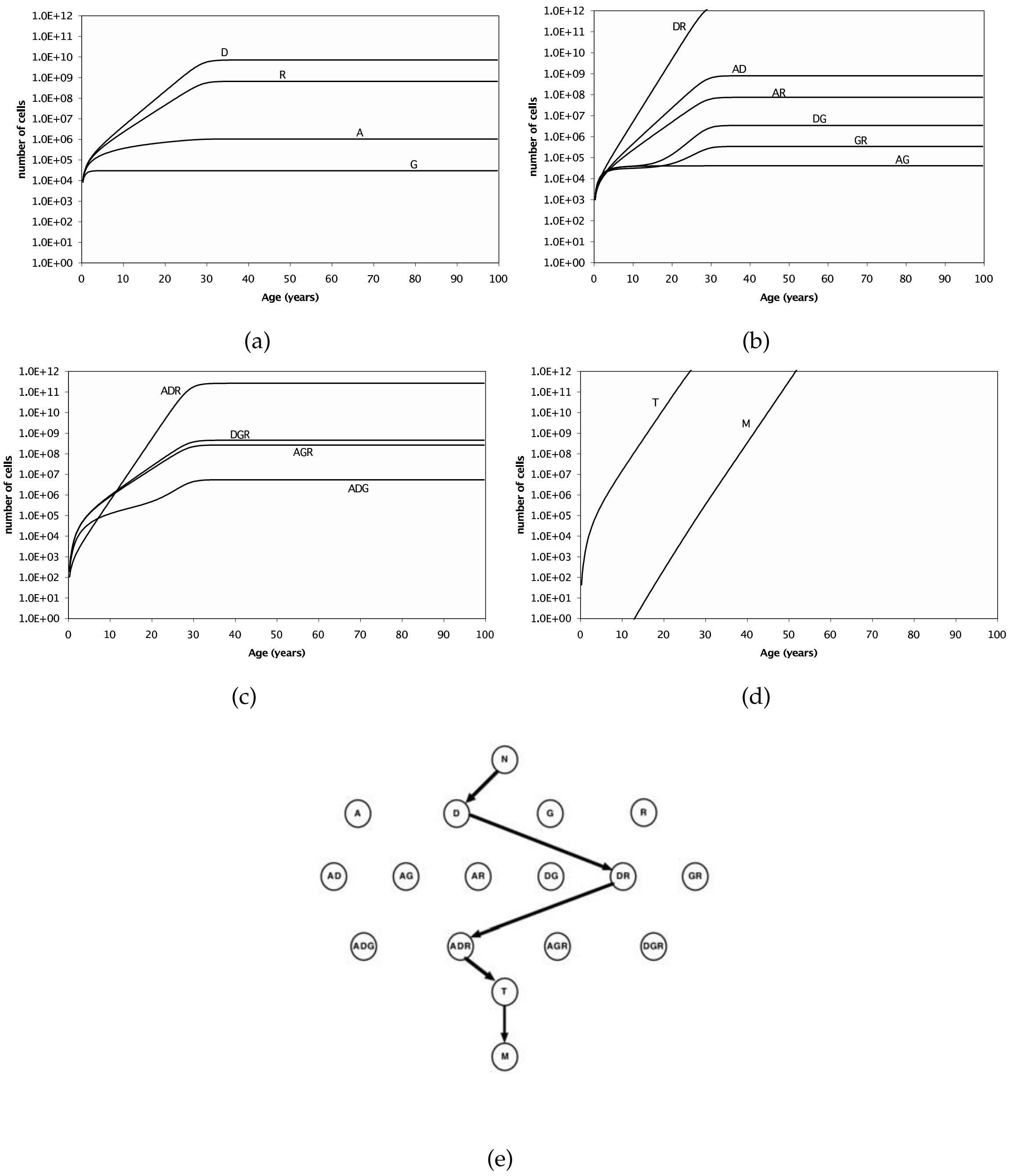
where T_R is tumour volume doubling time for cells with a mutation in R but not D and equals $T + 50 \times 3$, where T_D is the tumour volume doubling time for cells with a mutation in D but not R and equals $T + 50 \times 7$, and where the D:R importance ratio is 7:3. The base tumour volume doubling time is T when the growing tumour has mutations in both D and R (500 days). The value 50 is chosen to give realistic doubling times for cells with mutations in D (but not R) and R (but not D), on the upper bound of observed tumour volume doubling times, where the cells typically have both mutations.

7.5 Kinetics of various paths to cancer

Given that multiple mutations are necessary to form a tumour, we are interested in whether the specific order of mutations is important. It is currently believed that the temporal sequence of mutations determines the propensity of tumour development (Arends 2000). The extent to which genetic instability (G) determines the timing of tumourigenesis has been a controversial issue in cancer biology. Some have argued that an increased premalignant mutation rate (that is, acquiring a mutation in G early) is necessary for tumour development (Loeb 1991, Rajagopalan *et al.* 2003). Others have argued that an increased cell division rate, offering more opportunities to accumulate mutations, is sufficient for tumourigenesis (Tomlinson and Bodmer 1995, Tomlinson and Bodmer 1999, Sieber *et al.* 2003). Although the extent to which angiogenesis (A), decreased apoptosis (D), genetic instability (G), and increased replication rate (R) contribute to the development of cancer depends on the type of cancer involved, a better general understanding of the kinetics of various paths to cancer would be more informative about their relative importance.

The kinetics of various pathways to cancer were considered by analyzing the dynamics of the different cell populations. By plotting different sets of cell populations, one is able to identify the individual contribution of each mutation to the development of cancer. The growing populations of cells plateau at various points in the graphs due to the imposed 10^{13} cell population cap. In this model, the fastest pathway for tumour progression starts with a mutation in D, Figure 7.2(a), which increases the population of potential tumour cells. Next, a mutation in R is acquired, further increasing the population of cells by clonal expansion, Figure 7.2(b). After acquiring these two mutations, the tumour is sufficiently large to be inhibited by the angiogenesis cap imposed by the model. For this reason, a mutation in the angiogenesis category occurs next in the fastest path, Figure 7.2(c). Finally, a mutation in G follows. Figure 7.2(d) shows the populations of tumour cells, T, and metastatic cells, M. Although the rate $k_3 = 10^{-9}$ is very low, the large increase in population of T cells guarantees that eventually some cells successfully metastasise.

The model predicts that genetic instability is more likely to be a feature of later-stage sporadic tumours, in accordance with the view of Tomlinson and Bodmer (1999). This is because a mutation in G has no direct selective advantage, only an indirect advantage through increasing the mutation rates in other genes. Although genetic instability can aid tumourigenesis, selection and clonal expansion are the main driving force for tumour progression in this model, a conclusion which has been proposed previously by Sieber *et al.* (2003).



(a)

(b)

(c)

(d)

(e)

Figure 7.2. Fastest path to cancer. (a) Dynamics of cell populations with one type of mutation. (b) Dynamics of cell populations with two types of mutations. (c) Dynamics of cell populations with three types of mutations. (d) Dynamics of cell populations with four types of mutations (T), and those that have metastasised (M). (e) The fastest path to cancer is by acquiring a mutation in D, then R, then A, then G.

7.6 Effect of inherited mutations on cancer development

Here the effect of different inherited mutations on cancer development by varying the initial conditions was examined. Since most inherited cancers are the result of mutations in tumour suppressors (Knudson 2002), this situation is modelled by increasing the rate of transition from a normal cell to the appropriate mutated cell to 10^{-5} mutations / gene / cell division. This models a case where a person inherits an inactivating mutation in one copy of the gene. These cells are still functionally "normal" (thus they begin in state N), but the chance of acquiring the second "hit" and losing functionality of the protein (moving into the mutated state) is much increased.

As expected, inheriting a mutation in a cancer-critical gene decreases the time to cancer onset. The effects of inheriting a mutation in each category on time to reach 10^9 primary tumour cells, 10^{12} primary tumour cells, and 10^{12} metastatic cells are shown in Figure 7.3. A tumour volume of 1 cubic centimeter weighs about 1 gram and represents about 10^9 cells (Friberg and Mattson 1997). This tumour size is regarded as relatively small in a clinical setting and it is at this size that a tumour may give rise to the first symptoms and may first become detectable by palpation (Friberg and Mattson 1997), that is, by being physically felt by a physician. A tumour that weighs about 1 kilogram (10^{12} cells) is approaching the lethal tumour burden for a patient (Friberg and Mattson 1997). The 10^{12} metastatic cells plotted in Figure 7.3 are not necessarily localised to one site in the body; they could represent 10^{12} cells present in one location or 10^{11} cells present in each of 10 different locations, for example.

In contrast to the results obtained in Figure 7.2 where the increased population of cells caused by mutations in D and R dominates the fastest path to sporadic cancer, *inheriting* a mutation in a G gene causes cancer onset at the earliest age. There is no observable difference between inheriting a mutation in one of the other categories (A, D, or R) and inheriting no mutations at all (N), partly due to the robustness of the model to changes in parameters, discussed in the Figure 7.3 caption and in Section 7.7. In the fastest path plots (Figure 7.2), there is equal probability of acquiring a mutation in A, D, G or R. D will dominate over G due to the fact that the transition from one state to another is a function not only of the mutation rates k_1 and k_2 but also the cell population size. Both D and G are equally likely to begin with, but since D increases the net cell population very quickly, it soon dominates over the rate k_2 associated with G. Therefore, the fastest path to sporadic cancer is through a mutation in D first.

In comparison, when a mutation in G is *inherited*, the cell has already surpassed the initial probability hurdle of acquiring a mutation in G. The rate of subsequent mutation is now 1000-fold higher and once a mutation in D or R is obtained, the cell population will begin to increase. For this reason, an inherited mutation in G has the greatest effect. This result is consistent with the stochastic model of Nowak *et al.* (2004). The result is also consistent with the fact that many inherited cancer syndromes are the result of a mutation in the G category. These include xeroderma pigmentosum, ataxia telangiectasia, Nijmegen breakage syndrome, hereditary non-polyposis colorectal cancer, and Bloom syndrome (Sieber *et al.* 2003).

The time to develop a palpable primary tumour (10^9 cells) in this model is 16.25 years if no mutations are inherited (Figure 7.3). Even taking into account the fact that detection of the tumour would not occur until several years later (Friberg and Mattson 1997), this age of cancer onset is significantly earlier than the average age of cancer onset in the human population (Depinho 2000). This is an indication of the need for more accurate information on cell division, cell death, and tumour doubling rates. Importantly, this may also be an indication that acquisition of mutations in more than four categories is necessary for development of a primary tumour. Adding two more steps to the multistep model would certainly delay the time to cancer, however to add those as detailed in (Hanahan and Weinberg 2000) would be better left for an agent-based model. This has been considered by Spencer *et al.* (2006). Consideration of the role of the immune system in curbing the growth of a tumour would also slow the time to cancer onset. The model here does not directly consider this factor, although category D does allow for apoptosis initiated by the immune system. Consideration of these three factors would allow the model to be more appropriately scaled to the timing of human cancer.

Although it is somewhat inelegant to consider the time to onset of cancer in terms of time to reach a particular level using a deterministic model, the intent of the model was to consider the mean time to reach clinically significant numbers of cancer cells. The model could be extended to use stochastic differential equations. The use of an ordinary differential approach allows us to quickly explore the parameter space, which is done in the following section.

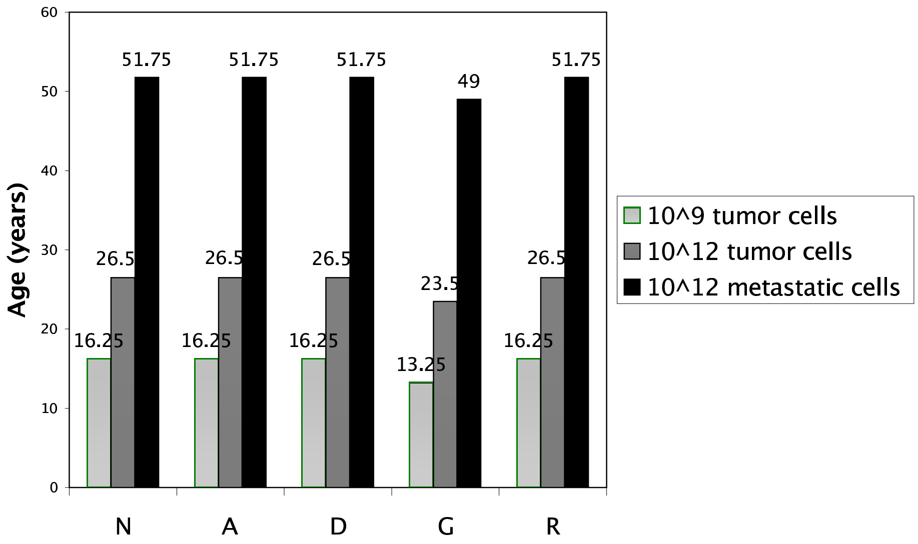


Figure 7.3. Inherited mutations in cancer-critical genes. Age at which a person may acquire 10^9 primary tumour cells, 10^{12} primary tumour cells, and 10^{12} metastatic cells with different inherited mutations. For reference, the case where no mutations are inherited is also shown (N). An inherited mutation in G is the only case which produces an earlier onset of cancer than the case where no mutations are inherited. Inheriting a mutation in G has a large effect because it allows the cell to surpass the initial probability hurdle of acquiring a mutation in G which then confers a 1000-fold increase in the rate of subsequent mutation. Increasing the initial probability of getting into state A, D, or R (that is, inheriting a mutation in A, D, or R) does not increase the time to cancer onset due to the number of cells that normally already build up in states D and R and due to the fact that mutations in A confer no benefit until 10^6 cells are obtained.

7.7 Sensitivity analysis of variations in the parameters

In order to determine the relative contributions of the parameters to the model, each parameter in Table 7.2 was varied while holding all others constant at the default value. The default values chosen are a best estimate from the literature. Except for the "D:R importance ratio" where 3:7 was used to determine the effect on the fastest path, and "% A cells needed to remove cap" where the range tested was from 0% to 100%, the other values were chosen to be near the upper and lower bounds of the range given in the literature. The contribution of each set of parameters was examined by examining their effect on time to reach 10^{12} M cells and on the fastest pathway to cancer, and the results are given in Table 7.2.

The most salient result of the sensitivity analysis is the robustness of the model. Despite trials with very high values for k_2, the fastest path to somatic cancer is always via a mutation in D then R then A then G, except in the case where the D:R importance ratio is flipped. As expected, a ratio of 3:7 flips the roles of D and R in the fastest path to give $RDAG$, but does not change the time to 10^{12} M cells from the default value of 51.75 years. Using a D:R ratio of 8:2 decreases the time to reach 10^{12} M cells due to the increased weight given to D.

The parameter that has the largest effect on time to reach 10^{12} M cells is the tumour volume doubling time. A tumour volume doubling time of 300 days decreases the time to reach 10^{12} M cells by 13.50 years relative to the default of 500 days, and a tumour volume doubling time of 700 days increases the time to reach 10^{12} M cells by 13.00 years. This effect is seen in Figure 7.4(a) as well as in Table 7.2. The large effect of this parameter on the model is due to its impact on the $\left(\frac{1}{b} - \frac{1}{d}\right)$ term; when cells have mutations in R and/or D, the terms become $1/b_R$ and/or $1/d_D$, allowing the cell populations to increase at a rate that reflects the tumour volume doubling time chosen.

Variations in the birth and death rates to 1 every 5 days and 1 every 30 days also have an effect on time to reach 10^{12} M cells. This can be seen in row one of Table 7.2, but the effect is small when compared with the effect of tumour volume doubling time.

The time (51.75 years) to reach 10^{12} M cells does not change in varying the percentage of A cells needed to remove the angiogenesis cap from 0% to 31%. Between 31% and 35%, the time to reach 10^{12} M cells increases rapidly. The time (57.50 years) to reach 10^{12} M cells does not change in varying the percentage from 35% to 100%. This effect can be seen in Figure 7.4(b). At 31%, the requirement for mutations in the A category begins to have an effect on the growing cell populations. At 35% and above, the percentage of A cells required is so large that the sum of the populations of non-normal, non-metastatic cells never goes above 10^6 because there are never at least 35% with A mutations. Thus the time to reach 10^{12} M cells depends only on a fixed number of T cells in each case, and remains constant at 57.50 years for percentages 35% and above.

A ten-fold change (from 10^{-7} to 10^{-6} mutations/gene/cell division) in mutation rate without a G mutation (k_1) has a larger effect on time to reach 10^{12} M cells than a ten-fold change (from 10^{-4} to 10^{-3} mutations / gene / cell division) in mutation rate with a G mutation (k_2). This is due to the fact that the effect of k_2 only becomes important later in tumourigenesis since G is last in the fastest path to cancer, whereas the effect of k_1 occurs at the beginning. There is no change in time to reach 10^{12} M cells when k_2 is increased from 10^{-3} to 10^{-2} mutations / gene / cell division because the effect of mutation rate has already saturated the system at a k_2 value of 10^{-3}.

Table 7.2. Sensitivity of the cancer model to changes in changes in parameters. Cell birth and death rates have units days^{-1}. Tumour volume doubling times are measured in days. Mutation rates are measured as mutations / gene / cell division. Number of genes involved in transitions are listed as number involved in single, double, triple transitions. "Other" refers to different values tested in the sensitivity analysis. "Time" refers to age at acquisition of 10^{12} M cells for a variation in that parameter, measured in years. "Path" refers to the fastest path to cancer for a variation in that parameter. $\mathcal{S}_1$ denotes the number of genes involved in transitions is $\{100, 10, 1\}$ for the single, double, and triple mutations; $\mathcal{S}_2$ denotes $\{500, 100, 10\}$. The blank boxes in the bottom right indicate only one alternative value was tried.

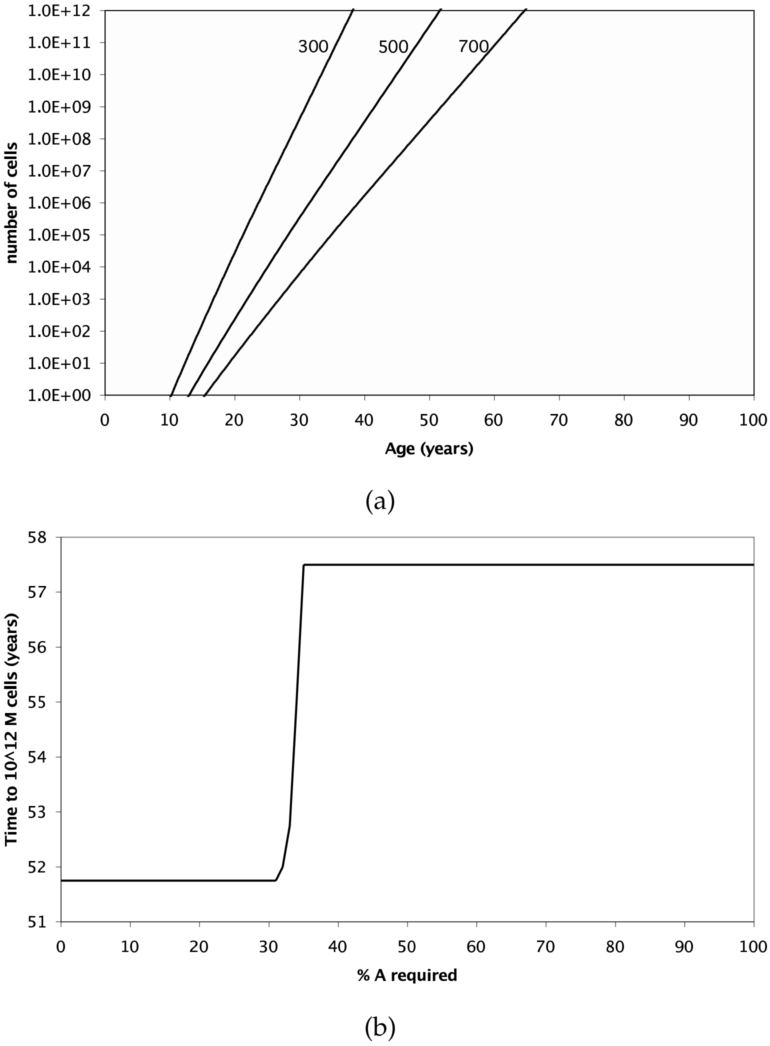
Parameter	Default	Time	Path	Other	Time	Path	Other	Time	Path
Cell birth and death rates	1/10	51.75	DRAG	1/5	50.25	DRAG	1/30	54.75	DRAG
Tumour volume doubling time	500	51.75	DRAG	300	38.25	DRAG	700	64.75	DRAG
% A cells needed to remove cap	10%	51.75	DRAG	30%	51.75	DRAG	40%	57.50	DRAG
D:R importance ratio	7:3	51.75	DRAG	8:2	50.50	DRAG	3:7	51.75	RDAG
Mut. rate with a G mutation	10^{-4}	51.75	DRAG	10^{-3}	50.50	DRAG	10^{-2}	50.50	DRAG
Mut. rate without a G mutation	10^{-7}	51.75	DRAG	10^{-6}	48.50	DRAG			
# of genes involved in transitions	$\mathcal{S}_1$	51.75	DRAG	$\mathcal{S}_2$	48.25	DRAG			

Since the parameter "number of genes involved in transitions" is located in the numerator of the differential equations, increasing the number of genes involved in the transitions decreases time to reach 10^{12} M cells simply by making the numerator larger.

To determine whether or not a more realistic age of cancer onset could be obtained, the model was run using all parameter values that would push back the time to cancer onset, but that are still in the biologically valid range. The parameters that were varied from the default settings are a tumour volume doubling time of 700 days, a cell birth and death rate of one every 30 days, and an angiogensis cap removal percentage of 40%. Running the model with these parameter adjustments results in a time to 10^9 tumour cells of 20.75 years, a time to 10^{12} tumour cells of 33.75 years, and a time to 10^{12} metastatic cells of 65.5 years. Although this time to cancer onset is more realistic, it does not represent the best estimates of parameter values from the current literature.

7.8 Conclusions

This chapter explores facets of the multistep model of oncogenesis. The key findings of this work are:



(a)

(b)

Figure 7.4. Sensitivity of the cancer model to changes in parameters. (a) Sensitivity to changes in tumour volume doubling time. Time to reach 10^{12} M cells is shown for tumour volume doubling times of 300 days, 500 days, and 700 days. (b) Effect of variations in percentage of A cells required to induce angiogenesis on time to reach 10^{12} M cells.

1. The fastest path to somatic cancer is predicted to be through gaining mutations in D (evasion of cell death), then R (increased replication rate), then A (angiogenesis), then G (increased mutation rate).

2. Of the four categories of mutations, inheriting a mutation in G is predicted to produce cancer at the earliest age, in line with known cancer epidemiology and other models of cancer progression (Nowak *et al.* 2004).

3. The fastest path to somatic cancer is robust to realistic changes in parameters, with the model being most affected by variations in tumour volume doubling time.

The strength of the model lies not in its utility for predicting any one individual's time to cancer onset *per se*, but rather in the fact that it presents a novel approach to understanding the genetic basis of cancer from a systems biology perspective. Although a thorough testing of this model is not currently possible due to the "generic" nature of the model, in that specific types of cancer and specific genes are not considered, this model establishes the groundwork for future models that can be directly tied to clinical and molecular data pertaining to a specific type of cancer.

It is hoped that the creation of this model for the multistep progression to cancer will encourage biologists to gather quantitative data and will suggest which experiments should be performed with highest priority. The only parameter values which are reasonably agreed upon in the literature are the spontaneous mutation rate and the size to which a tumour can grow before angiogenesis is required. All other parameter values could use experimental refinement. In particular, experimental data that would help include direct measurement of tumour volume doubling time and measurement of the number of mutations in various gene categories in heterogeneous cancer cell populations. The experimental data on mutations should consider the ordering of acquisition of key genes in each of the categories and include a study of inherited mutations in these key genes. A number of genes that could be categorised and tracked include those in the Wellcome Trust Sanger Institute database (Futreal *et al.* 2004). Tracking mutations in A category genes along with a study of the timing of angiogenesis would allow an estimate of the percentage of A cells required to signal for successful angiogenesis and tumour growth past 10^6 cells. Estimates of cell division and death rates in cells with these mutations would also be useful. The model is kept as general as possible; to make use of much of this experimental data the model would need to focus on a particular type of cancer, as many of these parameters are highly dependent on the originating tissue type.

To push back the time to cancer onset, a six step ODE model more like the one proposed by Hanahan and Weinberg was constructed, with the same four gene categories as the ODE model (A, D, G, R) and two additional ones: L (limitless replicative potential, e.g. turning on telomerase) and M (invasion and metastasis, e.g. loss of E-cadherin). The main problem was that L or M could not accurately modelled using ODEs. To properly model the effect of a beneficial mutation in L, one would need to store how many times each individual cell divides. A population-based ODE model ignores individual cells and their cell divisions, and only aggregate cell populations could be modelled. To properly model invasion, one would need to consider the cells existing in 3D space. The ODEs can only model change in population over time, not through space, and thus it was not possible to create equations that could reflect invasion by just using a rate change.

Better estimates of parameter values, inclusion of two additional categories to give a total of six steps in the multistep model, and consideration of the role of the immune system in curbing the growth of a tumour will allow future models to be more appropriately scaled to human cancers. Modelling the multistep accumulation of genetic mutations in cancer will give insight into topical questions about the progression of a normal cell to a cancerous cell, enabling cancer treatments to be better targeted to various stages of cancer progression, and suggesting the most important directions for future experimental research.

To summarise the novel contributions of this work: a mathematical model of the progression to cancer was developed and used to explore the pathways to cancer, the effect of various biological parameters, the average age at which cancer develops, and the affect of inherited mutations on this age.

Chapter 8

The human brain during sleep

Electroencephalograph (EEG) analysis enables the neuronal behaviour of a section of the brain to be examined. If the behaviour is nonlinear then nonlinear analysis tools can be used to glean information on brain behaviour, and aid in the diagnosis of sleep abnormalities such as obstructive sleep apnea syndrome (OSAS). In this chapter the sleep EEGs of a set of normal and mild OSAS children are evaluated for nonlinear behaviour.

Noise is present in the wide variety of signals obtained from sleeping patients. This noise comes from a number of sources: from presence of extraneous signals to adjustments in signal amplification and shot noise in the circuits used for data collection. The noise needs to be removed in order to maximize the information gained about the patient using both manual and automatic analysis of the signals. In this chapter a number of new techniques for removal of that noise are explored, along with techniques for the associated problem of separating the original signal sources.

8.1 Introduction

This chapter deals with two important issues in the analysis of EEG files:

1. There is a high degree of systematic and measurement noise present in recorded EEG signals.

2. If nonlinear tools are to be used, it first needs to be established if the EEG signals show significant nonlinearities. That is, can it be established if there is a possibility that a nonlinear model can generate the observed brain signal? In many papers, such as Fell *et al.* (2002) and Hwa and Ferree (2001), this is completely ignored or done very poorly. This chapter shows that while there are some sections of sleep EEGs that are significantly nonlinear, up to 40% of the time, the sleeping brain is not producing any signals that are significantly nonlinear! This has strong implications for the research being done, showing that traditional tools that assume linearity (or near-linearity) such as the Fast Fourier Transform are definitely valuable.

A linear time series is one where a model can be built that predicts the future samples based on the past samples (in general, in n-dimensional space), that is, there is a linear function

$$\boldsymbol{F}\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_{t-\tau}\right) = A\boldsymbol{x}_t + B\boldsymbol{x}_{t-1} + \ldots + T\boldsymbol{x}_{t-\tau} + \boldsymbol{\eta} = \boldsymbol{x}_{t+1} \tag{8.1}$$

where $A, \ldots, Z$ are $n \times n$ matrices for samples $\boldsymbol{x}$ (column vectors) in n-dimensional space (n channels of input) and $\eta \in \mathbb{R}^n$ represents random measurement noise, independent on the n channels of input at each time step. If a time series can be assumed to be linear, or close to linear, then it is amenable to linear methods that exploit the linear relationships, for example autocorrelation functions or related Fourier transform methods. If for example, the time series can be represented by a model that contains nonlinear terms like $\boldsymbol{x}^T A\boldsymbol{x}$ then it is said that the time series is nonlinear, and there are a number of nonlinear time series analysis techniques that could then be used. A couple of important points can be made:

- Reduction of the noise term $\boldsymbol{\eta}$ is important in both model construction and time series analysis.

- If it can be said that the model is a nonlinear model, then it suggests that the underlying dynamics of the system being modelled are also described by nonlinear equations.

It is for the latter point that establishing nonlinear behaviour (or its absence) in EEG time series is indicative of underlying nonlinear brain behaviour (or absence thereof), and the first point is the reason why a large part of this chapter is on noise removal techniques.

8.1.1 Novel contributions

The novel contributions of this work are:

1. Development of a methodology to firstly estimate the original, noisy, biological sources (in sleep patients) that have been mixed into several signals, then to clean them.
2. Analysis of the linearity (or nonlinearity) of brain systems as observed in their EEG signals, with the finding that a large portion of sleep does not show any significant nonlinearities.

8.2 Noise removal

8.2.1 Introduction

Electroencephalograph (EEG) and electrooculograph (EOG) measurement techniques provide valuable information on sleep disorders (Kaeming *et al.* 2003, Sforza *et al.* 2002). Recent studies looking at memory and learning during sleep have used these techniques as predictors of waking performance (Kaeming *et al.* 2003, Muzur *et al.* 2002, Anderer *et al.* 2002). Comparison of thoracic and abdominal movements associated with breathing can reveal important information about breathing disorders and events such as apneas and hypopneas during sleep (Brown *et al.* 2002, Prisk *et al.* 2002, Menon and Agrawal 2003).

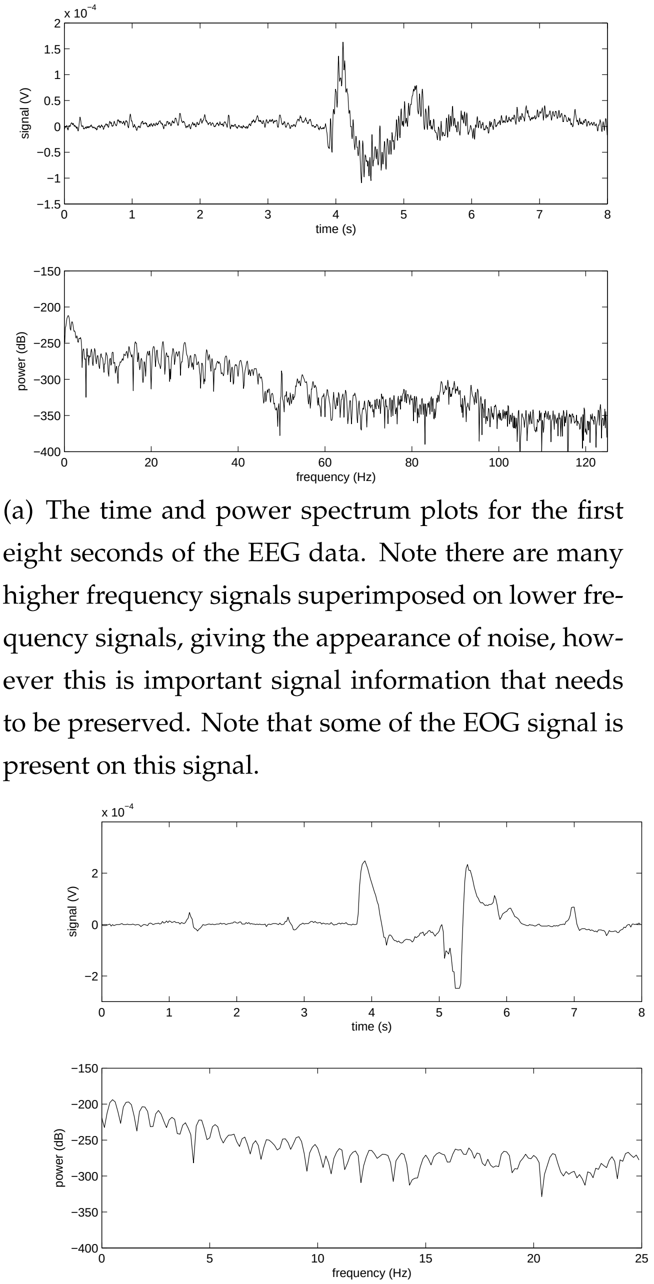
The process by which the EEG and EOG signals are recorded is described by Teplan (2002), but a brief summary is given here. The EEG and EOG signals are recorded by placing electrodes on the patient's scalp. These detect electric potentials generated by the flow of ions in neural cells that set up electric dipoles between the body of the neuron (soma) and the neural branches (apical dendrites). For the data used, these signals were amplified, then digitised at 250 Hz for the EEG and 50 Hz for the EOG signal, using a signed 8 bit digital format. The mutual information between the second EEG channel with the left EOG channel is estimated. The second EEG channel is measured from the left anterior position E1 to just below the opposite ear, with the left EOG channel measured from just to the side of the left eye to the position just above the nose between the eyes. The signals are broken down into (typically) 30 second long epochs, the combined set of signals is then classified by a human operator into various stages of sleep and wakefulness.

The main problems with analyzing the EEG and EOG signal are:

- Notch filtering of the the 50 Hz interference ripple (caused by power supply interference) from the signals also removes useful information.
- Skin conductances can vary over time in different ways in different locations, however the gel used to stick electrodes to scalp locations helps prevent this problem.
- Due to conductances across the skin, the signal received by an electrode is a mixture of the true signals one is trying to measure.

The most significant problem is the mixing of signals, which can be reduced by blind signal separation techniques using higher order statistics (Gorodnitsky and Belouchrani 2001, Belouchrani *et al.* 1997). The noise can then be removed using wavelet transforms (Matalgah and Knopp 1994, Bertrand *et al.* 1994, Lim *et al.* 1995). These techniques are also considered for the thoracic and abdominal movements, for which similar problems may arise (Menon and Agrawal 2003).

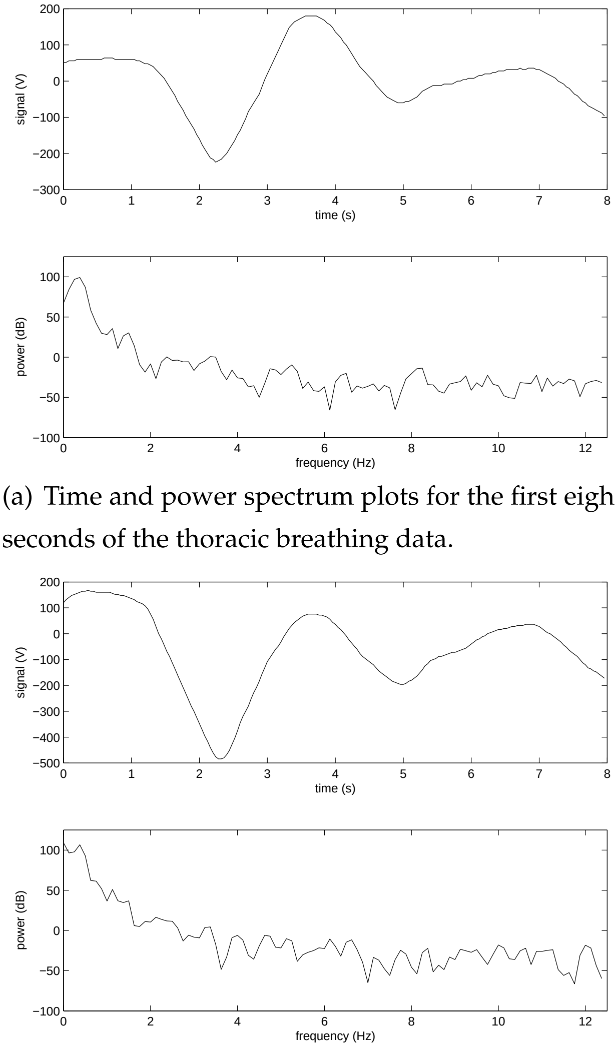
Time and power spectra plots for one of the EOG and EEG sets of data is shown in Figure 8.1. Those for one of the thoracic and abdominal sets of data are shown in Figure 8.2.



(a) The time and power spectrum plots for the first eight seconds of the EEG data. Note there are many higher frequency signals superimposed on lower frequency signals, giving the appearance of noise, however this is important signal information that needs to be preserved. Note that some of the EOG signal is present on this signal.

(b) The time and power spectrum plots for the first eight seconds of the EOG data. Note the lack of high frequency signals present in the EEG signal, since here we are concerned with low frequency muscle movement signals. The sampling rate used was correspondingly lower (50 Hz as opposed to 250 Hz for the EEG).

Figure 8.1. Time and power spectra plots for eye and brain wave data. The time and power spectra plots for the first eight seconds of the EEG and EOG data. Note the spectral differences between the two, with the EEG having many higher frequency components.



(a) Time and power spectrum plots for the first eight seconds of the thoracic breathing data.

(b) Time and power spectrum plots for the first eight seconds of the abdominal breathing data.

Figure 8.2. Time and power spectra plots for thoracic and abdominal breathing data. The sampling rate was 25 Hz, allowing the capture of relatively slow breathing signals. Note the thoracic signal has a slight phase lead over the abdominal breathing signal.

8.2.2 Methods

There are several problems to solve in eliminating noise from the signals. For the problem of the EOG and EEG signals, one must first ensure the EOG and EEG signals have the same number of data points. To do this a Gaussian smoothing procedure can be used, that is detailed in Subsection 8.2.3. While other related smoothing procedures could be used, here one can assume the distribution of the signal data is Gaussian—which is reasonably true for the data used—and thus a Gaussian smoothing procedure could be used. The problem of separating the sources from the observed signals, which contained a mixture of both, can be dealt with. Three algorithms are evaluated for this, detailed in Subsections 8.2.6 to 8.2.8. The noise was removed from both of the sources using wavelet transforms as elaborated on in Subsection 8.2.9. A flowchart of the process is shown in Figure 8.3.

To evaluate the performance of the blind signal separation used to separate the source data, and to evaluate the noise removal, an efficient algorithm, given in Subsection 8.2.10, was used to estimate the mutual information between two signals. One would expect this measure to decrease when comparing the original signals with the separated signals, assuming greater independence between the separated signals, and to remain the same when comparing the noisy signals with those where the noise has been removed, assuming the noise is uncorrelated between the two signals. This is largely true for the signals of interest, although there may be some information at certain frequencies that is correlated due to extraneous electromagnetic signals being received by the leads, as they act as antennas. This is kept to a minimum through appropriate grounding.

8.2.3 Gaussian smoothing

The EEG signal has a sampling rate five times higher than the EOG (250 Hz to 50 Hz). These are recorded simultaneously, so every fifth time point in the EEG corresponds to a time point of the EOG signal. Gaussian smoothing is used to reduce the number of data points in the EEG by a factor of five:

$$s\left(i\right) = \frac{\sum\limits_{j=i-m}^{i+m} x_j w\left(x_j, x_i\right)}{\sum\limits_{j=i-m}^{i+m} w\left(x_j, x_i\right)}, \tag{8.2}$$

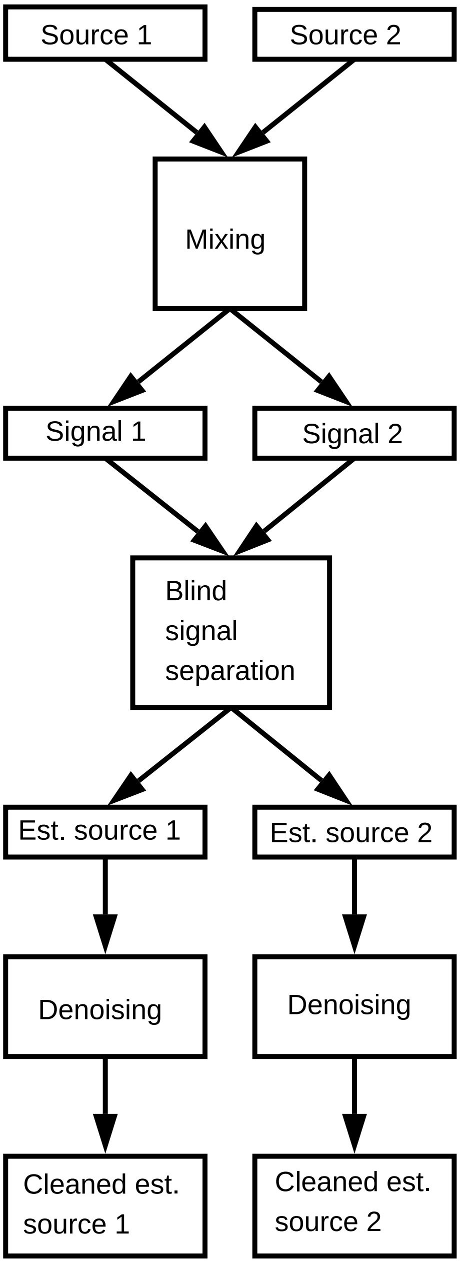


Figure 8.3. Flowchart of signal processing steps. Flowchart showing the general steps involved in going from the original sources to the cleaned, estimated sources. The optional Gaussian smoothing step is not shown.

where the weights $w\left(x_j, x_i\right)$ are

$$w\left(x_j, x_i\right) = e^{-\left(x_i - x_j\right)^2/\left(2\hat{\sigma}^2\right)}, \tag{8.3}$$

i is the discrete time point the smoothed average is being calculated for, and $\hat{\sigma}^2$ is the estimate of variance for the entire set of samples in the EEG signal. To the two sets of data, which can be written as $\boldsymbol{x}\left(t\right) = \left[x_1\left(t\right), x_2\left(t\right)\right]^T$, one can apply blind signal separation, using the following model of the data.

8.2.4 Data model for blind signal separation

One can write the original, m-dimensional source data as $\boldsymbol{s}\left(t\right) = \left[s_1\left(t\right), s_2\left(t,\right), \ldots, s_m\left(t\right)\right]^T$. It is assumed that the sources are independent. One then considers an unknown linear model $\boldsymbol{A}_{n\times m}$ as generating the observed signals, where $\boldsymbol{A}_{n\times m}$ is written as an n-dimensional vector $\boldsymbol{x}\left(t\right) = \left[x_1\left(t\right), x_2\left(t\right) \ldots, x_n\left(t\right)\right]^T$ by

$$\boldsymbol{x}\left(t\right) = \boldsymbol{A}\boldsymbol{s}\left(t\right), \tag{8.4}$$

where $\boldsymbol{A}$ is referred to as the mixing matrix. Note that an arbitrary swap of columns of $\boldsymbol{A}$, and scaling a source by a scaling change in a row of $\boldsymbol{A}$, means there is an ambiguity in both the permutation (of labeling) of the sources and the scaling of the sources respectively. With this model of the data, one can then apply blind signal separation techniques.

8.2.5 Blind signal separation

There are two key blind signal separation approaches that are combined to form the joint cumulant and correlation (JCC) algorithm in Subsection 8.2.8. They are the second order blind identification (SOBI) algorithm, discussed in Subsection 8.2.6, and the joint approximate decomposition of eigenmatrices (JADE), discussed in Subsection 8.2.7. Both approaches have a common first step, in which the data is whitened using a *sphering* matrix $\boldsymbol{W}$, which transforms the mixing matrix $\boldsymbol{A}$ into a unitary matrix $\boldsymbol{U}$, which is a matrix for which $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$ (Gorodnitsky and Belouchrani 2001, Belouchrani *et al.* 1997). The next step of estimating $\boldsymbol{A}$ is dependent on the choice of algorithm and detailed below.

8.2.6 SOBI algorithm

Given a hypothesis of sources with different spectra and the linear model of Eq. 8.4, one can calculate time-delayed, cross-correlation matrices,

$$\begin{aligned} \boldsymbol{R}(\tau) &= \mathrm{E}\left[\boldsymbol{x}(t)\,\boldsymbol{x}(t-\tau)^T\right] \\ &= \boldsymbol{A}\boldsymbol{R}_s(\tau)\,\boldsymbol{A}^H, \end{aligned} \tag{8.5}$$

where $\tau \neq 0$ and

$$\boldsymbol{R}_s = \begin{pmatrix} \mathrm{E}\left[s_1(t)\,s_1(t-\tau)\right] & 0 & 0 & \cdots & 0 \\ 0 & \mathrm{E}\left[s_2(t)\,s_2(t-\tau)\right] & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & & \vdots \\ 0 & \ldots & & 0 & \mathrm{E}\left[s_m(t)\,(t-\tau)\right] \end{pmatrix}, \tag{8.6}$$

with $E[\cdot]$ the expectation operator. The correlation matrices can then be whitened,

$$\underline{\boldsymbol{R}} = \boldsymbol{W}\boldsymbol{R}(\tau) = \boldsymbol{U}\boldsymbol{R}_s(\tau)\,\boldsymbol{U}^T, \tag{8.7}$$

$\forall t \neq 0$. The joint diagonalization of the set of p whitened correlation matrices $\{\underline{\boldsymbol{R}}(\tau_i)\,|i = 1, \ldots, p\}$ gives the matrix $\boldsymbol{U}$ (Gorodnitsky and Belouchrani 2001). The matrix $\boldsymbol{U}$ can only be uniquely determined iff for any (i, j), there exists at least one lag τ_k such that $\mathrm{E}\left[s_i(t)\,s_i(t-\tau)\right] \neq \mathrm{E}\left[s_j(t)\,s_j(t-\tau)\right]$ (Gorodnitsky and Belouchrani 2001). The mixing matrix is then estimated by $\hat{\boldsymbol{A}} = \boldsymbol{W}\boldsymbol{U}$. An alternative to the SOBI algorithm is the JADE algorithm.

8.2.7 JADE algorithm

Here the linear model of Eq. 8.4 is assumed, as is independence of sources. To each n-dimensional vector x is associated a quadicovariance matrix $\boldsymbol{Q} : \boldsymbol{M} \rightarrow \boldsymbol{N}$ defined by $\boldsymbol{N} = \boldsymbol{Q}\boldsymbol{M}$ such that

$$N_{i,j} = \sum_{(k,l)} \mathrm{Cum}\left(x_i, x_j, x_k, x_l\right) M_{k,l}, \tag{8.8}$$

where $\mathrm{Cum}\,(\cdot)$ is defined as

$$\mathrm{Cum}\left(x_i, x_j, x_k, x_l\right) = \mathrm{E}\left[\bar{x}_i\bar{x}_j\bar{x}_k\bar{x}_l\right] - \mathrm{E}\left[\bar{x}_i\bar{x}_j\right]\mathrm{E}\left[\bar{x}_k\bar{x}_l\right] - \mathrm{E}\left[\bar{x}_i\bar{x}_k\right]\mathrm{E}\left[\bar{x}_j\bar{x}_l\right] - \mathrm{E}\left[\bar{x}_i\bar{x}_l\right]\mathrm{E}\left[\bar{x}_j\bar{x}_k\right], \tag{8.9}$$

and where $\bar{x}_i = x_i - \mathrm{E}\left[x_i\right]$, etcetera (Cardoso 1999). As the set of $n \times n$ matrices is an n^2-dimensional linear space, it can be shown that there exist n^2 real eigenvalues λ_r and n^2

orthonormal eigenmatrices $\boldsymbol{M}_r$ satisfying $\boldsymbol{Q}\boldsymbol{M}_r = \lambda_r \boldsymbol{M}_r$ (Gorodnitsky and Belouchrani 2001). It can be proved that only n of the eigenvalues are non-zero (Cardoso and Souloumiac 1993), and that joint diagonalizaton of the n corresponding eigenmatrices, labelled $\underline{\boldsymbol{M}}_r$, gives the unitary matrix $\boldsymbol{U}$ (Gorodnitsky and Belouchrani 2001). As with the SOBI algorithm, the mixing matrix is estimated by $\hat{\boldsymbol{A}} = \boldsymbol{W}\boldsymbol{U}$. Combining the two algorithms gives us the JCC algorithm.

8.2.8 JCC algorithm

In the JCC algorithm, both the correlation information provided by the SOBI algorithm of Subsection 8.2.6, $\underline{\boldsymbol{R}}(\tau_i)$, and the cumulant quadricovariance eigenmatrices, $\underline{\boldsymbol{M}}_r$, provided by the JADE algorithm, are used. Joint diagonalization gives the unitary matrix $\boldsymbol{U}$, which again acts to give an estimator $\hat{\boldsymbol{A}} = \boldsymbol{W}\boldsymbol{U}$. Using $\hat{\boldsymbol{A}}$, one can then separate the signals into estimates of the original source data. The noise can be removed from this data using wavelet techniques.

8.2.9 Wavelet noise removal

The mathematical description of the continuous wavelet transform (CWT) of $f \in \boldsymbol{L}^2(\mathbb{R})$ is described by Mallat (1999) as

$$(Wf)(u,s) = \int_{-\infty}^{+\infty} f(t)\,\psi_{u,s}^{*}(t)\,dt, \tag{8.10}$$

where

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\phi\left(\frac{t-u}{s}\right), \tag{8.11}$$

is a family of orthogonal wavelets, $\|\psi_{u,s}\| = 1$, $\langle \psi_{u,s}, \psi_{u',s'} \rangle = 0$ for $(u,s) \neq (u',s')$, and

$$\int_{-\infty}^{+\infty} \psi_{u,s}(t)\,dt = 0. \tag{8.12}$$

The scale of the wavelet may conceptually be considered the inverse of the frequency.

The CWT reveals much detail about a signal, however due to the continuous nature it cannot be computed for real signals on a digital computer. Therefore, the discrete wavelet transform (DWT) is normally used. The DWT calculates the wavelet coefficients at discrete intervals of time and scale instead of at all scales. With the DWT, a fast version of the algorithm is possible, analogous to the fast Fourier transform. This version of the algorithm makes use of the fact that if scales and positions are chosen based on powers of two (dyadic scales and positions) the analysis is very efficient. In 1988, Mallat developed an efficient way to implement this algorithm, which is known as a two-channel sub-band coder (Mallat 1989). For a single level of decomposition, this algorithm passes the signal through two complementary (high-pass and low-pass) filters resulting in approximations which are high-scale, low-frequency components of the signal, and details, which are low-scale, high-frequency components of the signal. This results in twice as many data-points so the data is down-sampled. For further levels of decomposition, successive approximations may be iteratively broken down into details and approximations as shown in Figure 8.4. Coefficients below a certain level are regarded as noise and thresholded out. Thresholding may be soft or hard. Hard thresholding is defined as

$$y = \begin{cases} x & \text{for } |x| > \theta, \\ 0 & \text{for } |x| \leq \theta, \end{cases} \tag{8.13}$$

and soft thresholding as

$$y = \begin{cases} \mathrm{sign}(x)(|x| - \theta) & \text{for } |x| > \theta, \\ 0 & \text{for } |x| \leq \theta, \end{cases} \tag{8.14}$$

where x is the original signal, y is the thresholded signal, and θ is the threshold. Hard thresholding tends to create discontinuities at $x = \pm\theta$ because any values of the signal less than the threshold are immediately set to zero. With soft thresholding, the thresholded values are shrunk towards zero without creating the discontinuities. The signal is then reconstructed without significant loss of information. Then the signal may be reconstructed by up-sampling, passing the approximations and details through the appropriate reconstruction filters and combining the results. Based on SNR measures of wavelet performance, Daubechies wavelets of order 5 were used, with soft thresholding and a decomposition level of 5; although this is not the best for noise removal, it is more important to preserve information when going from the estimated sources to the denoised estimated sources.

To evaluate the performance of the above techniques a particular measure of mutual information was used.

8.2.10 MI estimation algorithm

The following is an outline of the method used in calculating the mutual information between the EEG and EOG signals, as given in Kraskov *et al.* (2003).

The mutual information for two signals X and Y is defined in Eq. 8.15

$$I\left(X,Y\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mu\left(x,y\right) \log \frac{\mu\left(x,y\right)}{\mu_x\left(x\right)\mu_y\left(y\right)} dxdy, \tag{8.15}$$

where μ, μ_x and μ_y are probability measures. One then takes the set of points $z_i = (x_i, y_i)$ for the EEG x_i and EOG y_i, $i = 1, \ldots N$. Then one finds the kth closest neighbor of each z_i according to the metric

$$\|z - z'\| = \max\{\|x - x'\|, \|y - y'\|\}. \tag{8.16}$$

The kth nearest neighbor is then projected onto the x and y axes giving the distances $\epsilon_x\left(i\right)/2$ and $\epsilon_y\left(i\right)/2$ respectively. The mutual information is estimated by:

$$\hat{I}_k\left(X,Y\right) \approx \psi\left(k\right) - 1/k - \langle\psi\left(n_x\right) + \psi\left(n_y\right)\rangle + \psi\left(N\right), \tag{8.17}$$

where $\psi\left(\cdot\right)$ is the digamma function given by

$$\psi\left(z\right) = \frac{d}{dz}\ln\Gamma\left(z\right), \tag{8.18}$$

and

$$\langle\ldots\rangle = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\left[\ldots\left(i\right)\right]. \tag{8.19}$$

8.2.11 Results

Blind signal separation

The blind signal separation abilities of the three algorithms were tested, across four sets of data, two of thoracic and abdomen (TA) breathing data, and two of EEG and EOG (EE) data. The differences in mutual information between the signal data and the estimated source data are shown in Table 8.1. The higher the mutual information, the better the algorithm is for separating the original sources, given the assumptions of that algorithm.

Table 8.1. Mutual information differences between estimates sources and original signals. The difference in mutual information (in nats/sample) between the two estimated sources and the two signals, $\hat{I}_k$ (est. sources) $-$ $\hat{I}_k$ (signals), for the two sets of thoracic-abdominal (TA) data and the two sets of EEG and EOG (EE) data. Nats are units of information, when a natural logarithm is used. This is computed for all three (SOBI, JADE, and JCC) algorithms

	TA1	**TA2**	**EE1**	**EE2**
SOBI	1.0319	0.4539	0.4522	0.6101
JADE	< 0.0001	< 0.0001	< 0.0001	< 0.0001
JCC	0.3640	0.4292	0.4032	0.5342

Table 8.2. Mutual information differences between JADE estimated sources and wavelet denoised estimated sources. The difference in mutual information (in nats/sample) between the JADE estimated sources and the wavelet denoised estimated sources is computed for the two sets of thoracic-abdominal (TA) data and the two sets of EEG and EOG (EE) data, $\hat{I}_k$ (est. denoised sources) $-$ $\hat{I}_k$ (est. sources). Nats are units of information, when a natural logarithm is used.

TA1	TA2	EE1	EE2
< 0.0001	< 0.0001	< 0.0001	0.0001

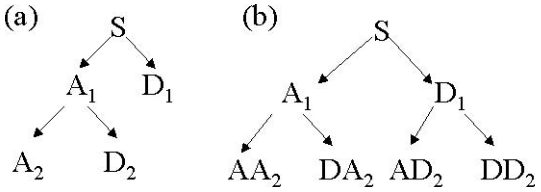


Figure 8.4. Wavelet decomposition. This figure illustrates (a) how the discrete wavelet transform decomposes a signal into details and approximations iteratively decomposing the approximations, and (b) how the wavelet packets iteratively decompose the approximations and details.

Wavelet denoising

For each of the generated estimates of the sources (three blind signal separation algorithms applied to four pairs of signals) the mutual information between the estimates of the sources was calculated, and the denoised estimates of the sources. These are given in Table 8.2. No difference was observed in mutual information between the estimates and denoised estimates, indicating little lost of information.

8.3 Nonlinear analysis

8.3.1 Overview

Electroencephalograph (EEG) analysis enables the neuronal behaviour of a section of the brain to be examined (Teplan 2002). As neurons themselves display nonlinear behaviour, it is suspected that the overall behaviour of groups of neurons is also nonlinear (Elbert *et al.* 1994). If the behaviour is nonlinear, it allows the use of nonlinear statistics to describe the behaviour of the brain (Kantz and Schreiber 1997).

Measurement and analysis of the EEG is an integral part of the evaluation of sleep disorders in both adults and children. It is used in the classification of sleep architecture, a cyclic progression of sleep that is tightly controlled such that in adults a new cycle of REM (rapid eye movement) and Non-REM sleep occurs approximately every 90 minutes. A common respiratory sleep disorder is obstructive sleep apnea syndrome (OSAS). In OSAS the upper airway experiences repetitive periods of partial or complete occlusion during sleep. The disruption of sleep architecture by OSAS leads to well described daytime sequelae including reduced neurocognitive functioning, increased problematic behaviour, daytime sleepiness, impaired mood, and an increased risk of accidents. EEG parameters in combination with respiratory data are used to assess OSAS severity and these have been correlated with deficits in daytime functioning. EEG parameters can be derived through linear and nonlinear analyses. Evidence of linear and nonlinear brain activity has been demonstrated in adults (Stepień 2002, Das and Das 2004) but very little research has been done in children; in particular there are conflicting results with different measures (Ferri *et al.* 2002, Ferri *et al.* 2003) and between children and young adults. It also remains to be demonstrated whether any observed nonlinearity reflects brain processes rather than nonlinearity of the amplifiers and other equipment used to collect the EEG data. Given the central role of sleep in neuronal development and plasticity it is imperative to establish in children the relationship of linearity and nonlinearity in brain behaviour during sleep. This may also provide novel insights into the mechanism and effects of sleep architecture disruption caused by OSAS. In particular, it is important to test whether nonlinear parameters distinguish normal children from those with OSAS.

8.3.2 Participants

Thirteen children with a history of snoring and suspicion of OSAS participated in this study. These children had been referred to a paediatric sleep disorders unit for evaluation of upper airway obstruction prior to adenotonsillectomy. In addition, 13 non-snoring controls of a similar age range were also recruited into this study from friends of the snoring group or through newspaper advertisements. All children underwent an overnight polysomnogram (PSG) to evaluate the degree of upper airway obstruction and to collect EEG data. Other than a history of snoring in the former group, all children were otherwise healthy and not taking any medication that may influence EEG dynamics. Informed consent was obtained from all parents of the children and, where age appropriate, from the children themselves. This study was approved by the South Australian Women's and Children's Hospital Research Ethics Board.

8.3.3 Overnight polysomnography

Overnight polysomnography (PSG) was conducted without sedation or sleep deprivation and began at each child's usual bedtime utilising standard protocols for children (American Thoracic Society, 1994). A parent accompanied each child throughout the procedure. The following standard parameters were measured and recorded continuously: electroencephalogram (EEG; C3-A2 or C4-A1), left and right electrooculogram (EOG), sub-mental and intercostal electromyogram (EMG) with skin surface electrodes, leg movements by piezoelectric motion detection, heart rate by electrocardiogram (ECG), oro-nasal airflow by thermistor and/or nasal pressure, respiratory movements of the chest and abdominal wall using uncalibrated respiratory inductive plethysmography (RIP), arterial oxygen saturation (SaO2) by pulse oximetry (three second averaging time) and transcutaneous CO2 (TcCO2) using a heated (314 K) transcutaneous electrode.

All polysomnograms were analysed and scored manually by a sleep technician experienced and trained in analysing paediatric sleep studies. Sleep stages were scored in 30-second epochs according to the standardised EEG, EOG and EMG criteria of Rechtschaffen and Kales (1968) and included rapid eye movement (REM) sleep and the four stages (1-4) of non-rapid eye movement (NREM) sleep. As stage 3 NREM sleep comprises only a small proportion of children's sleep it was combined with stage 4 NREM sleep and termed slow wave sleep (SWS) as is common practice. Respiratory variables were scored according to standard guidelines recommended for paediatric sleep studies (Marcus *et al.* 1992, Society 1996). Obstructive apnoeas were defined as the absence of airflow associated with continued chest and abdominal wall movement for a duration of two or more respiratory cycles. Obstructive hypopnoeas were defined as a 50-80% reduction in the amplitude of the RIP and/or airflow signal associated with paradoxical chest/abdominal wall movement for a duration two or more respiratory cycles associated with either a 4% oxygen desaturation and or EEG arousal.

8.3.4 EEG recordings

The EEG data was recorded from the C3-A2 or C4-A1 position in the international 10-20 electrode placement system, with a reference point behind the mastoid. The signal was notch filtered to remove as much of the 50 Hz AC ripple as possible and amplified by an analog amplifier. The analog signal was sampled at 125 Hz and digitised using a linear digitizer. Artifact contamination of the EEG signals included extraneous signals from muscular movement (Drinnan *et al.* 1996), digitization noise, and also signals from the environment being picked up by unshielded EEG leads. Of particular concern is the nonlinear nature of filters and amplifiers used to process the analogue signal before digitization, as what is of interest is the nonlinearities in the underlying brain processes and not those of the equipment used.

8.3.5 EEG data analysis

As discussed by Schreiber and Schmitz (1997), there are a number of methods for determining whether signals originate from nonlinear models or not.

There are a number of caveats with using these; many of them require assumptions about the data and have varying power of rejection of the null hypotheses of linearity. Of these, the best one overall seems to be the simple time reversibility test, which only requires assymetry of the data under visual inspection. The time reversibility test computes a simple time reversal statistic on the data under test, and a set of linear surrogate data chosen carefully with the same general statistical properties (Schreiber and Schmitz 1996). The particular statistic is given by

$$t_r = \frac{\sum\limits_{n=i_d+1}^{N} \left(x\left[n\right] - x\left[n - i_d\right]\right)^3}{\sum\limits_{n=i_d+1}^{N} \left(x\left[n\right] - x\left[n - i_d\right]\right)^2}, \tag{8.20}$$

where i_d is a delay. Typically $i_d = 1$ is sufficient, and is used in this work. The test works by computing a value based on powers of the sample points - if the system has a linear model, then this statistic will be significantly different (in fact less) than if the system had a nonlinear model, that is, if future values could be predicted using nonlinear powers of the past samples. As Schreiber and Schmitz (1997) note, this measure works best when there are only a few data sets with clear asymmetry under time reversal. Windows of length 10 000 from the EEG files that have significant end effects are used, and from this 19 sets of surrogate data of the same length are generated that have the same Fourier amplitudes and distribution; this provides a better null hypothesis than using a Gaussian linear process (Schreiber and Schmitz 1996). If the value of t_r for the original data is not the least out of the set of surrogate data, then the null hypothesis is rejected at the 95% significance level, and hence shows nonlinearity. The reason why end effects—or that the variance of the signal changes through the time window across which samples are taken—is important, is that this means that there is a greater likelihood of t_r being different in nonlinear samples, since the top line will increase much more than the bottom line would due to the different powers of an increasing signal variance (or magnitude).

It would be prudent to also use another measure of nonlinearity, and here the Higuchi fractal measure is used as it gives a number representative of the amount of nonlinearity in each individual window. The Higuchi fractal metric gives us a measure of the underlying nonlinear dynamics of a signal without trying to reconstruct a strange attractor (Kantz and Schreiber 1997, Higuchi 1988, Accardo *et al.* 1997). The Higuchi measure provides a reliable measure of the fractal dimension when working with short time series segments, that is, those with sample length $N < 125$ (Accardo *et al.* 1997). It is also relatively insensitive to nonlinearities in noise or in amplification (Accardo *et* *al.* 1997) so is useful for establishing that the nonlinear behaviour comes from the underlying system, in this case the brain. For the Higuchi fractal metric, one first calculates

$$L(k) = \sum_{m=0}^{k-1} \frac{N-1}{\lfloor \frac{N-m}{k} \rfloor k^2} \sum_{j=1}^{\lfloor \frac{N-m}{k} \rfloor} \left| x(m+jk) - x(m+(j-1)k) \right|, \tag{8.21}$$

where $N = f_s \times 8\text{s}$ for an eight second window (hence $N = 1000$ for $f_s = 125$ Hz, or $N = 2000$ for $f_s = 250$ Hz), and $k = 1, 2, \ldots, 2f_s$. Using a least squares fit of $y = \log(L(k))$ against $x = \log k$ gives the Higuchi fractal measure d_H,

$$d_H = -\frac{\operatorname{cov}(x,y)}{\operatorname{Var}(x)}. \tag{8.22}$$

The Higuchi fractal measure lies between 1 and 2 in theory; in practice because it is only an estimate it may lie slightly outside this range. The lower the value the “less complex” and linear the signal is. Higher values indicate signals that look more complicated and are nonlinear.

To compare results between the two groups and between sleep stages, the unpaired Student’s t-test is used, which first computes a t-value,

$$t = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{s_a^2}{N} + \frac{s_b^2}{M}}} \tag{8.23}$$

where $\bar{x}$ denotes the mean and s denotes the standard deviation for the two groups a and b with sizes N and M respectively. The t-value from Eq. 8.23 is then compared with a two-tailed Student’s t-distribution of $N + M - 2$ degrees of freedom to determine a significance level. This requires the data be approximately normal, and in the central limit theorem if there is enough data in both sets of data under consideration then the t-test can be safely used. This was checked, along with a check of the significance value by manually computing the probability distributions involved.

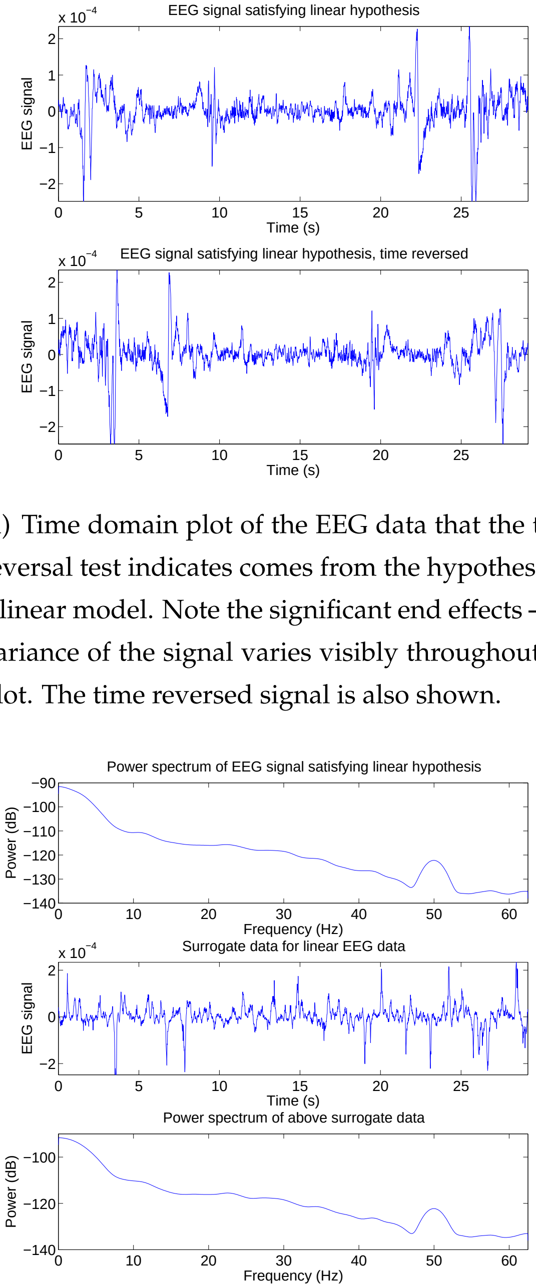
8.4 Time reversal results

8.4.1 Participants and PSG findings

The thirteen snoring children, six males and seven females, had a mean age of 6.8 years (range 5.1 – 8.7 years). The control group, also comprised of six males and seven females, had a mean age of 7.6 years (range 5.2 – 10.9 years). Overnight PSG analysis demonstrated that the snoring children had a higher number of obstructive apnoeas and hypopnoeas per hour of sleep (mean (± SD) 0.6 (0.90)) than the non snoring control group (mean (± SD) 0.03 (0.06)) and this difference was statistically different (P = 0.01, Mann Whitney U analysis). However as the number of obstructive breathing events was less than one per hour of sleep in the snoring children, this is considered as having only very mild OSAS, or snoring. There was no significant difference in the amount of time that each group of children slept (7.84 hours for the snoring group vs 7.17 hours for the control group). Similarly there was no significant difference in the amount of time spent in each sleep stage by both groups of children.

8.4.2 Verifying time reversal test

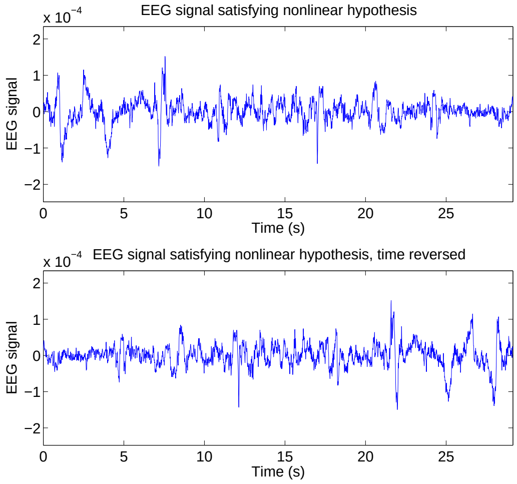
Visual verification was performed to check that the data has significant end effects, in order that the time reversal test can be used, and also that the surrogate data generated has the same power spectra as the original data. Figures 8.5 and 8.6 show that the data contains end effects, so the time reversal test can be safely used; furthermore they show that the surrogate method is correctly generating data with the same power spectra as the input data (the original time series) to the surrogate generation process.



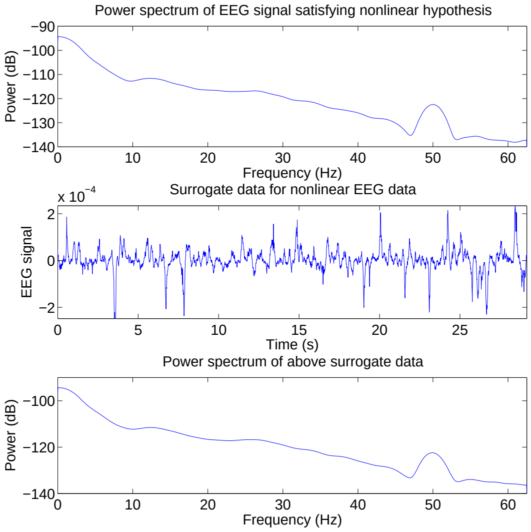
(a) Time domain plot of the EEG data that the time reversal test indicates comes from the hypothesis of a linear model. Note the significant end effects – the variance of the signal varies visibly throughout the plot. The time reversed signal is also shown.

(b) The power spectrum of the EEG data fitting the linear hypothesis as shown in Subfigure (a). Surrogate data generated by a linear model with the same power spectra as the linear EEG has been generated and is plotted along with its power spectra to check they are identical.

Figure 8.5. Time plots of EEG data fitting the linear hypothesis. The time plots of the EEG data fitting the linear hypothesis, its time reversal, and the surrogate data. Power spectra are also shown.



(a) Time domain plot of the EEG data that the time reversal test indicates comes from the hypothesis of a nonlinear model. Note the significant end effects – the variance of the signal varies visibly throughout the plot. The time reversed signal is also shown.



(b) The power spectrum of the EEG data fitting the nonlinear hypothesis as shown in Subfigure (a). Surrogate data generated by a linear model with the same power spectra as the nonlinear EEG has been generated and is plotted along with its power spectra to check they are identical.

Figure 8.6. Time plots of EEG data fitting the nonlinear hypothesis. The time plots of the EEG data fitting the nonlinear hypothesis, its time reversal, and the surrogate data. Power spectra are also shown.

Table 8.3. Mean percent of time the EEG shows significant behaviour. Mean percent of time (with standard deviation) the EEG was significantly nonlinear during different sleep/wake states computed from the time reversal statistic. Data are presented for clinical children, normal children, and for both groups combined.

Sleep stage	Patient group	Control group	Both groups
Non-REM 1	75.4 (12.7)	68.0 (16.3)	71.7 (14.8)
Non-REM 2	72.1 (8.7)	69.5 (9.1)	70.8 (8.8)
SWS	68.7 (14.1)	65.8 (12.6)	67.3 (13.2)
Wake	66.9 (9.9)	52.8 (11.3)	59.8 (12.7)
REM	74.8 (6.7)	67.5 (10.9)	71.1 (9.6)

Table 8.4. T-statistic comparing nonlinearity between sleep stages. Student's t-values comparing the amount of nonlinearity between sleep states, for the combined data set of control and mild OSAS children. Significance values: * = 95%, **=99%, ***=99.9%.

Sleep stage	REM	1 Non-REM	2 Non-REM	SWS
Wake	3.62***	3.10**	3.63***	2.07*
REM		0.163	-0.121	-1.21
Non-REM 1			0.259	1.14
Non-REM 2				1.14

8.4.3 Time reversal test results

For each sleep stage (including wake time after sleep onset), the percentage of the time is shown for which the time reversal test indicated significant nonlinear behaviour (at the 95% level of significance). The results are shown for all the children combined and for the control and mild OSAS group separately, in Table 8.3. Table 8.4 shows the differences (using the unpaired Student's t-test) in amount of nonlinear behaviour between different sleep states.

8.4.4 Higuchi fractal results

For each sleep stage (including wake after sleep onset) Higuchi fractal measures were calculated for all epochs in all subjects. These results are shown for the control and patient groups separately and for both groups combined in Table 8.5. Table 8.6 shows the differences (using the unpaired Student's t-test) for Higuchi fractal measures between different sleep states.

Table 8.5. Higuchi fract measure applied to EEG data. Higuchi fractal measures (mean ± SD) calculated for each 30 second epoch of data across the group of clinical children, the group of normal children, and both groups combined. The Higuchi fractal measure here indicates that the data is generally linear across all sleep stages and groups, including periods of waking between periods of sleep. The values for the patient group are typically higher, although this is not significant.

Sleep stage	Patient group	Control group	Both groups
Non-REM 1	1.113 (0.084)	1.079 (0.095)	1.099 (0.090)
Non-REM 2	1.113 (0.052)	1.097 (0.069)	1.105 (0.061)
SWS	1.099 (0.038)	1.067 (0.081)	1.083 (0.066)
Wake	1.103 (0.140)	1.095 (0.095)	1.098 (0.116)
REM	1.127 (0.041)	1.115 (0.061)	1.121 (0.051)

Table 8.6. T-statistic comparing Higuchi fractal measure between sleep states. This table shows the Student's t-values for sets of Higuchi fractal measures between sleep states, with the total set of data from both (control and patient) groups. Significance values: * = 95%, **=99%, ***=99.9%

Sleep stage	REM	1 Non-REM	2 Non-REM	SWS
Waking	13.3***	0.200	4.53***	-8.47***
REM		-10.4***	-17.2***	-35.3***
1 Non-REM			-2.60**	6.53***
2 Non-REM				21.9***

8.4.5 Discussion

For the particular set of thoracic and abdominal breathing data used, the SOBI algorithm works well, with an increase in the mutual information, probably because the sources have reasonably distinct spectra. Since the JCC combines information from both the SOBI and JADE algorithms by way of joint diagonalization, it introduces the problems associated with using the JADE algorithm for this data, namely that the sources are not independent. The two sources have a high level of dependence, being almost synchronous during regular breathing, tending to differ only for compliant chests in young children or when a breathing obstruction occurs (Brown *et al.* 2002, Menon and Agrawal 2003). Similarly, for the EEG and EOG data, although these are more independent, the SOBI algorithm performs best at separating the original sources from the observed signals.

The wavelet denoising performs well, in that it preserves (as far as was determined) the information present in the signals. Further work will consider wavelet packet and matching pursuit denoising algorithms (Mallat 1989, Mallat and Zhang 1993, Krishnan and Rangayyan 2000), and how these effect mutual information between two different channels. One could also consider the effect of swapping the denoising and blind signal separation techniques. In theory this should have little to no difference on the results.

Comparing the same sleep states between patient and control groups reveals no significant difference in percent nonlinearity for all sleep states. Comparing Higuchi fractal measures however, a significant difference between patient and control groups was found, but in REM sleep only (t-value 9.45). This sleep state has been shown to be associated with learning (Huber *et al.* 2004, Ficca and Salzarulo 2004), and disruptions to this sleep, as occurs in OSAS children, affects learning (Drummond *et al.* 2000).

The typically low values of the Higuchi fractal measure, being close to one, confirm the general linear trend indicated by the time reversal test. It also reveals that the nonlinearity is due to underlying brain behaviour and not instrument noise. The Higuchi fractal measure is insensitive to ergodic noise. The amount of nonlinear brain behaviour is highest in NonREM stages 1 and 2 in addition to REM sleep, and lowest during wake (after sleep onset) and slow wave sleep. This is not a surprising finding for slow wave sleep, with relatively predictable, low frequency waveforms present. It is somewhat surprising for wake after sleep onset, however it may represent simply the presence of linear muscle signals, transmitted across the skin, contaminating the EEG signal. More advanced techniques than those discussed could be used to remove this noise.

Given that this work has established nonlinearity in sleep stages of interest, in particular those associated with memory, it would make sense to use nonlinear measures to try and capture the brain behaviour. Linear measures (such as the often-used Fourier transform) should not be discounted however, since there is clear linearity throughout all sleep stages, and the signal may still be considered to be relatively stationary over local regions even when nonlinearity is present, as indicated by the low Higuchi fractal measures. Using the Higuchi fractal measure reveals a significant difference between control and patient groups in REM sleep, and this will be explored further in future work. It remains to be seen whether nonlinear measures are useful in classifying sleep stages. This work has highlighted the fact that the Higuchi fractal measure does not appear useful for classification in children, who present difficulties even for highly trained technicians in classifying sleep stages.

8.4.6 Conclusions

The work presented in this chapter has established some nonlinearity in the processes generating the EEG data using the time reversal and Higuchi tests, however there are considerable amounts of data that do not appear to be generated by nonlinear processes, in line with Stepień (2002). Due to the significant changes in nonlinearity between sleep stages, and the Higuchi fractal measure, one can be certain that the nonlinearity process arises in the brain and not as a result of any nonlinear processes in the recording equipment. In the process of determining this, a method was developed for cleaning up noise from sources including the recording equipment

This work highlights the need to test for nonlinearity before using nonlinear measures in evaluating EEG measure, in particular in distinguishing different brain states. Future work should focus on both linear and nonlinear measures for detecting local sleep events, such as apneas, as these may affect memory consolidation during sleep. The Higuchi fractal measure may be useful for this.

To summarise the novel contributions of this work:

- a methodology for cleaning up biological (sleep) signals was developed, and
- it was shown that there is a large amount of sleep (in children) that is significantly nonlinear, and that this amount varies with the scored sleep stage.

Chapter 9

Metabolomics

METabolomics is about performing data analysis on samples from complex systems such as cells, organs, and whole human beings, and is thus clearly in the field of complex systems research. In this chapter, two problems are considered: firstly distinguishing cancer cells grown in cell culture media, and secondly, on distinguishing urine samples from autistic children and their non-autistic siblings. Although these are noisy, complicated sources of biological data, and the mass spectroscopy can add further noise and systematic biases to the samples, some statistical separation between groups of samples can be shown to exist, and a statistical model was successfully built to distinguish between groups of samples.

9.1 Introduction

To repeat the definition from Chapter One, metabolomics is the study of metabolic output of biological systems to analyse their inner workings. In this chapter, some methods are developed for tackling these two main questions using metabolomics:

- Can autistic children be distinguished from their matched, non-autistic siblings, using only the metabolic output in their urine?
- Can different types of cancer cells be distinguished from each other, using their metabolic output as captured in the growth media that they are grown in?

Both sets of samples were run through a quadrupole liquid chromatography / mass spectroscopy unit. Basically this breaks molecules into fragments, ionizes them, and then they pass through sets of metal plates that have an electric field between them. Depending on the use of these electric fields, various things can be done based on the atomic mass to electronic charge ratio; in the case of these experiments, the quadrupoles are used to restrict the range of mass/charge ratios analysed, and to scan over this range for each retention time, or time taken for the sample to pass through the system. This gives a set of intensities: a count of how many particles hit the detector, over a scan of mass / charge ratios, which takes an interval of approximately one second of retention times. This is a very data-rich set of samples, that needs various steps to clean up the data, and then some processing steps to distinguish firstly if there are statistically significant differences, and then attempt to use these in grouping and classifying samples.

To use the data, one first needs to do some pre-processing steps to clean the data and then use basic statistics to assess if it contains any useful information. Then a number of statistical methods can be applied to glean information about, and use, statistical relationships between samples. This list includes (unsupervised) k-means clustering analysis, principal components analysis (PCA), and support vector machines (SVMs).

9.1.1 Novel contributions

The novel contributions of this work are:

1. Development of an algorithm that cleans, normalises, and then processes (using existing machine learning algorithms) mass spectroscopy data into a variety of

forms to show differences between groups of data and to build a model to recognise these differences.

2. Use of the t-statistic as a distance measure to feed into the neighbour-joining algorithm, as opposed to a statistical measure of the significance of the branch lengths (Pinto *et al.* 2003, Susko *et al.* 2002).

3. Use of machine learning techniques to identify unclassified urine samples as either from autistic or non-autistic children.

4. Use of machine learning techniques to identify unclassified cell metabolite samples as from either Huh7, HepG2, or HeLa cell lines.

9.2 Autism study participants

A group of children with autism and their matched siblings was recruited. At the closing date, 22 autistic children and 22 matched, non-autistic siblings had been successfully recruited. To be used in the study, these needed to meet careful criteria for both autism and also to establish the non-autistic controls. Ethical clearance was obtained as detailed below (subsection 9.2.2). Samples were collected appropriately—a minimum of two, first-void samples per patient, to assess intra-individual variation—and stored at appropriate temperatures and thawed before use. There were a number of procedures in place to ensure that samples were labelled and processed without any mixup.

9.2.1 Inclusion and exclusion criteria

The austim inclusion criteria were: children who have received a diagnosis of autism using standard diagnostic tools by a multi-disciplinary team and who are on the waiting list for or who are enrolled in the South Australian Early Intervention Research Program (EIRP) for children with autism. To find a group of matched non-autistic siblings, siblings of autistic children were chosen, making sure to exclude those siblings who have:

- deficits in communication, socialisation and/or stereotypic/repetitive behaviours,
- intellectual disability,

- one or more co-morbidities (for example, epilepsy), or
- a chromosomal abnormality.

9.2.2 Ethical clearance

The autism research was carried out with the authorisation of both the University of South Australia Ethics Committee and The Flinders University of South Australia Ethics Committee. All children with autism who enrolled in or are on the waiting list of the Early Intervention Research Program (EIRP) for children with autism at Flinders University were screened for potential study participants.

9.2.3 Cancer cells used

Huh7 and HepG2 liver cell lines, and the HeLa cervical cancer line were grown in standard growth media, grown to 50% confluence, then the media replaced and the cells grown to 80% confluence. The second round of growth media was run through an Applied Biosystems API 3000 LC/MS/MS (liquid chromatography and dual mass spectroscopy) machine. Standard degradation of the growth media was controlled for by including control data where the growth media was set without cells present, and also run through the mass spectroscopy machine. Statistical testing revealed significant differences (p=0.01) between all three cell lines; in particular there were generally more differences between the HeLa cervical cancer cell line growth media and growth media with liver cell lines growing in them.

9.3 Methods

9.3.1 Pre-processing and preliminary analysis

Before data were analysed, all LC-MS data were first standardised for intensity $I(x,t)$ (count of particles per second) at various mass/charge ratios $\{x\}$ and retention times $\{t\}$. Note that although each scan over mass / charge ratios takes a segment of retention times, this segment is of fixed duration and only the starting time of each scan is given in the set $\{t\}$.

The two (or three) spectral data of each subject were averaged (some subjects may have three spectral data because one of their urine samples was analysed twice for intraday-variation monitoring). The data were analysed both with and without a total ion count step. Normalizing the data by the total ion count aims to offset differences due to variation in dilution volume. The final spectra then had the square root of each intensity component taken, having verified that the data comes from a roughly Poisson point process, in line with Purohit and Rocke (2003). Various PR approaches were then adopted in an attempt to identify any patterns existing in the spectral data and to classify samples based on the spectral properties.

Students paired t-test (Kullback 1968) was used to determine if there was significant variation between each set of LC-MS data. Given two sets of paired data, of values measured at $n = |\{x\}||\{t\}|$ points $I(x,t)$, the paired t-test determines whether they are significantly different, under the assumption that the paired differences are independent and the data are normally distributed (hence, the Central Limit Theorem applies (Feller 1945, Trotter 1959)). Then define

$$\hat{x}_i = x_i - \bar{x}, \tag{9.1}$$

$$\hat{y}_i = y_i - \bar{y}, \tag{9.2}$$

where the x_i and y_i are the paired samples $I_1(x,t)$ and $I_2(x,t)$ for a particular x and t, and $\bar{x}$ and $\bar{y}$ the respective means. The t-value was calculated using the formula

$$t = (\bar{x} - \bar{y})\sqrt{\frac{n(n-1)}{\sum_{i=1}^{n}(\hat{x}_i - \hat{y}_i)}}. \tag{9.3}$$

The absolute value of the calculated t-value was then used to represent graphically the relationships between samples or groups of samples using the neighbour-joining algorithm (Saitou and Nei 1987, Studier and Keppler 1988). The neighbour-joining algorithm describes these relationships using a distance which was in this case the absolute value of the t-value. A larger distance between samples or groups of samples signifies greater differences. This would hopefully allow us to establish if there were any significant general differences between samples. Dendograms represent graphically the relationships between samples, and are generated by the neighbour-joining algorithm using Student's paired t-test as a statistical distance between samples.

9.3.2 k-means clustering analysis

The k-means clustering method was used to classify samples together into k groups based on the spectral properties (Bishop 1995, Jain and Dubes 1990, Kaufman and Rousseeuw 1990). The grouping was done by minimizing the sum of squares of distances between data and the corresponding cluster centroid, in line with Kanungo *et al.* (2002). The raw data (in this case just the set of $I(x,t)$ values of the urine spectra) and the number of clusters, k ($k = 2$ as the samples comprised of both autistic and non-autistic categories), were input into the computer algorithm, which then attempted to group the samples into k clusters. This was to hopefully separate out the autistic and non-autistic samples from one another, and also the types of cancer cells. The k clusters were subsequently compared with the actual assignment of samples to each respective category to determine the accuracy rate.

Note that k-means clustering analysis has several limitations that may render successful identification of clusters difficult in the context of this study. First, to identify two clusters, the representative sample for each cluster will have to be predefined. As it is impossible to tell the differences in attributes of each sample, one assumes that each attribute has the same weight and thus the extent to which it contributes to the grouping process is neither known nor predictable. Secondly, selecting a different sample may affect the location of the centroid and hence the outcome of the clusters. Thirdly, the algorithm is highly sensitive to outliers thus resulting in possible deviations from the shape of the true cluster to accommodate data too distant from the true centroid. Given the small sample size of this study, even the order in which the data is fed into the computer may produce different clusters. Therefore, the k-means clustering algorithm may not be sufficiently robust, and thus lack the sensitivity in detecting true clusters from the perspective of this study.

9.3.3 Principal components analysis

Principal components analysis (PCA), in line with Yeung and Ruzzo (2001), was used to extract latent biochemical information, or principal components (PCs) from the complex spectral data sets. This would, hopefully, facilitate the visualisation of the patterns that lie in the data by segregating out the samples and displaying the sample groups in different clusters in either a 2-dimensional or 3-dimensional PCA scores plot.

9.3.4 Support vector machines

Support Vector Machines (SVM) are one form of machine learning (Christianini and Taylor 2000, Burges 1998) that is used to classify some set of data into two or more categories based on a learned model of the data from a training set consisting of data vectors $\boldsymbol{x}$ (one for each sample) with a known label (or category) y. The model is then tested on a different test set over which the model predicts a set of labels, which are then compared with known labels. Of interest is the success rate of prediction of the model. The goal of SVM learning is to find the optimal boundary that separates the clusters of vectors (in this case, each vector is formed from the set of intensity values $I(x,t)$ for all mass/charge (m/z) values $\{x\}$ and all retention times $\{t\}$ for each individual sample) in such a way that features of one group of samples are on one side of the boundary (or hyper-plane) and features of a different group of data are on the other side. The vectors near the boundary are known as the support vectors and these are used to construct the boundary, which is then formulated as a classifier function. Figure 9.1 helps illustrate this further. Appendix B provides more information on the optimisation mathematics behind construction of this classifier function. In most cases, however, the vectors are not easily separated by a linear plane. Instead, they may be better represented by a nonlinear surface (in general, these are n-dimensional problems with a high n). The beauty of SVM is that such complex boundaries can still be constructed by using the same algorithm with a different kernel function. It can take on many mapping functions with some of the most common types being the linear, polynomial, radial basis function (RBF), and sigmoid types. Therefore, the kernel function allows SVM to perform separations even in the presence of complex boundaries.

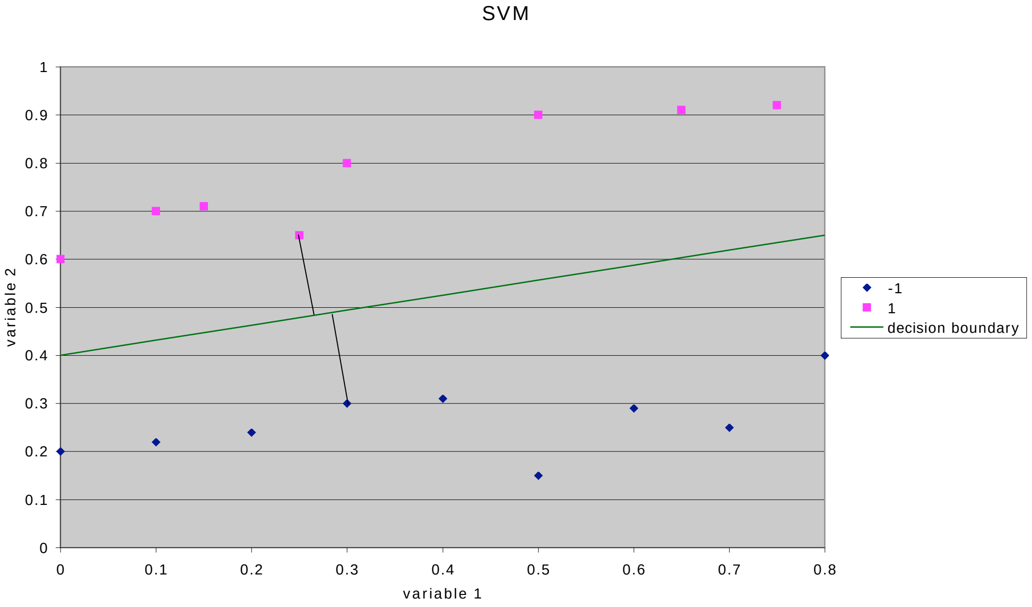


Figure 9.1. Sample SVM boundary. SVMs construct a boundary or hyper-plane that separates two classes (this can be extended to multi-class problems) in a 3D space. The hyper-plane is oriented so that the margin between the support vectors is maximised and points misclassified is minimised.

9.4 Cancer cell results

9.4.1 Pre-processing and preliminary analysis

The following algorithm was used to establish significant differences between the samples:

1. Firstly, normalize each individual run by the total ion count, controlling for differences in injection volume in the mass spectroscopy machine.
2. Establish no significant intra-run variation.
3. Integrate over a set of 30 second retention time intervals to control for peak shifts with retention time, in line with Jonsson *et al.* (2004).
4. Take the square root of the data, having verified it is roughly a Poisson point process, and in line with Purohit and Rocke (2003).
5. Compute the set of absolute values of the differences between the means of each category, and test if any are significant using the threshold test below. Ignore those that are significantly different between the controls.

Comparing the t-value with Students t-distribution gives a significance value. The Kolmogorov-Smirnoff test (Conover 1971) allows us to check if the intensity values come from a Poisson point process. There were no significant differences ($p = 0.05$) between the runs, once they were normalized by the total ion count. The Kolmogorov-Smirnoff test (Conover 1971) and stem plots were used to verify the intensity values come from a roughly ($p = 0.1$) Poisson point process.

Setting a threshold at the mean + 3 standard deviations (of the set of absolute differences between the sets of data) identifies those differences at the 99.7% significance level, assuming the differences come from a normal distribution (Kullback 1968). The central limit theorem applies, however it is still important to check the significance level, which was found to be $p \approx 0.01$. The formula for the threshold is given by:

$$\theta = \frac{1}{n}\sum_{i=1}^{n} x_i + 3\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{9.4}$$

where the x_i are the set of absolute differences across the range of mass/charge ratios. Table 9.1 shows the significant differences ($p = 0.01$) between the pairs of cell line

Table 9.1. Mass/charge ratios at which differences are significant. This shows mass/charge ratios for which the difference between the means of each group is significant (p=0.01), excluding those differences caused by general media degradation. This list was calculated for a variety of retention time intervals, summing up the ion count within that interval. Generally, HeLA shows up as being quite distinct from Huh7 and HepG2 (in particular HepG2).

Retention times	HuH7, HepG2	Huh7, HeLa	HepG2, HeLa
0-30 s	38.4, 44.9, 53.2	36.8, 37.8, 44.9	36.8, 37.0, 37.5, 37.6, 37.8
30-60 s	(none)	(none)	36.9, 37.4, 37.5
60-90 s	36.5, 36.8	36.5-37.3 (inclusive)	36.6, 36.7, 36.8, 36.9
90-120 s	(none)	(none)	36.9,37.0
120-150 s	40.0, 52.1, 46.8	38.6	41.1
150-180 s	46.4	54.9	57.1, 58.7
180-210 s	55.9	(none)	(none)

growth media, for 30 second intervals of retention time. Figure 9.2 shows a typical spectrum obtained for the Huh7 cells, and Figure 9.3 shows a comparison between spectra of samples taken at different time points.

The t-statistic is then used as a distance that can be fed into the neighbour-joining algorithm. This produces the graphs shown in Figures 9.5 and 9.4. Apologies for the overlap, it is difficult to avoid using the otherwise excellent *phylip* software package. The results indicate a good clustering on the non-normalised data, which is somewhat surprising given that the t-statistic and neighbour-joining algorithm do not implicitly handle these differences. However, it is clear there must be some base-line differences in the samples that are occuring. This is unlikely to be due to some sort of systematic error, as even the controls cluster together, and repeated experiments on the same types of cells also cluster together.

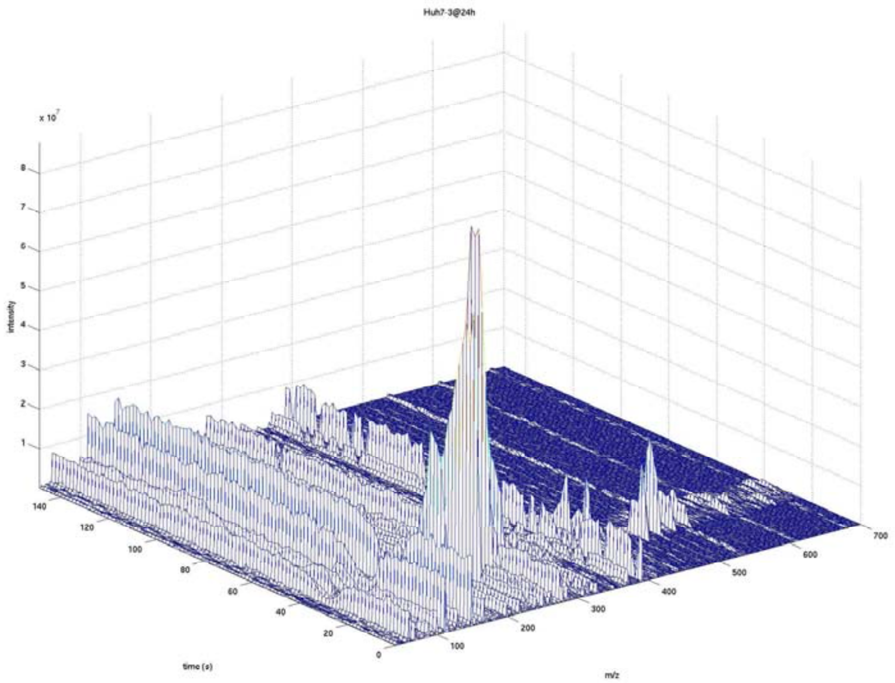
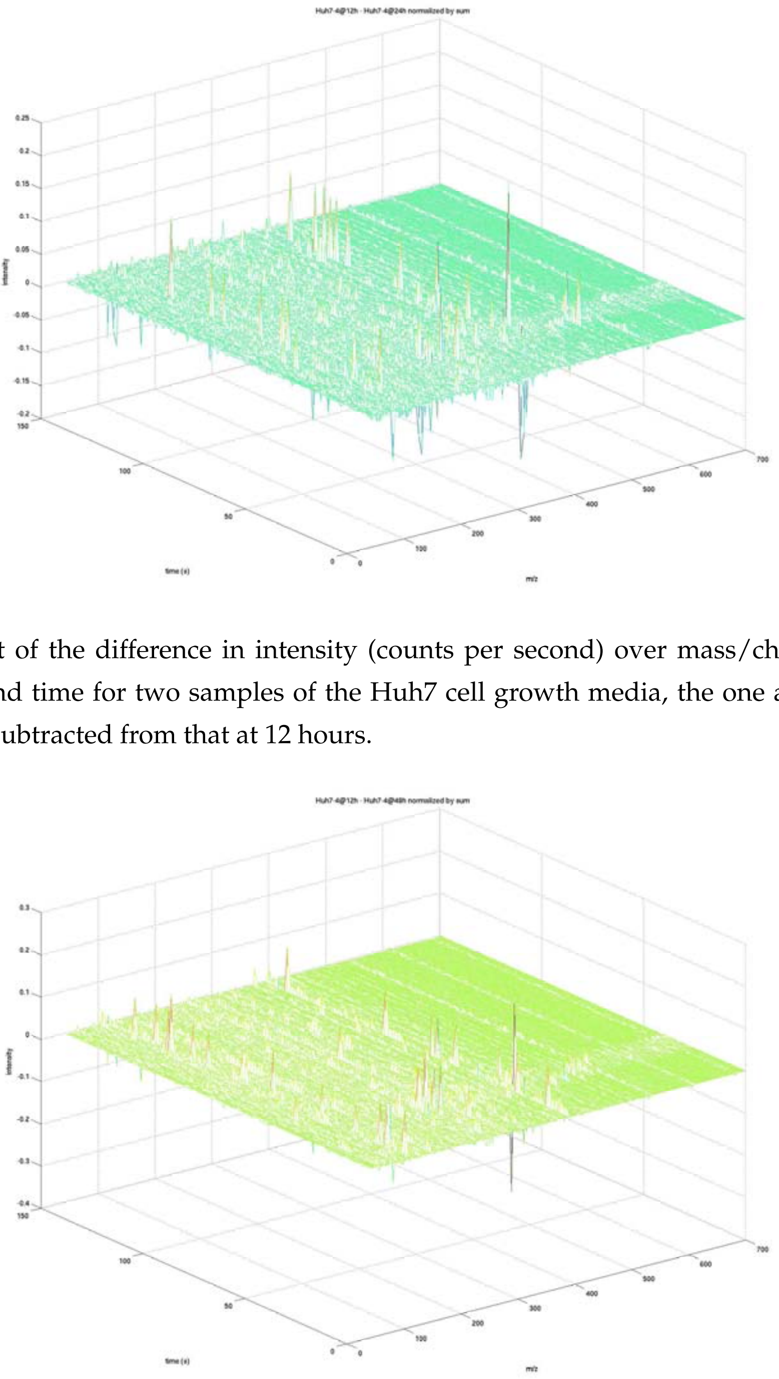


Figure 9.2. Mass spectral plot of Huh7 cell growth medium. Plot of the intensity on the z, or vertical axis (in counts per second), over mass/charge ratio (y, or right hand side axis) and time (x, or left hand side access) for a sample of the Huh7 cell growth media. The time at sampling was 24 hours after the Huh7 cells started growing.



(a) Plot of the difference in intensity (counts per second) over mass/charge ratio and time for two samples of the Huh7 cell growth media, the one at 24 hours subtracted from that at 12 hours.

(b) Plot of the difference in intensity (counts per second) over mass/charge ratio and time for two samples of the Huh7 cell growth media, the one at 48 hours subtracted from that at 12 hours.

Figure 9.3. Mass spectra plots of changes in growth media over time. Plot of the difference in intensity on the z, or vertical axis (in counts per second), over mass/charge ratio (y, or right hand side axis) and time (x, or left hand side access) for samples of the Huh7 cell growth media at three time points.

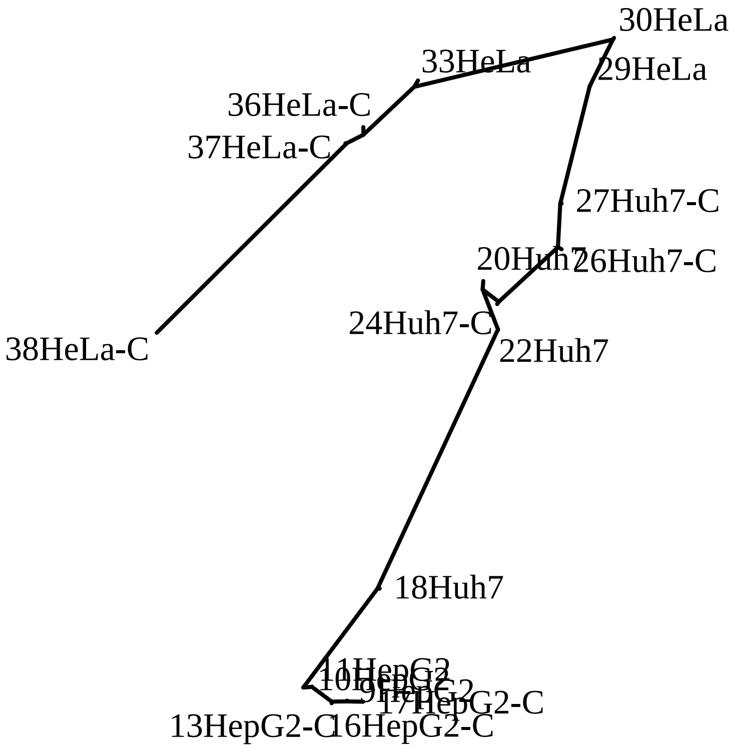


Figure 9.4. Grouping of non-normalised cancer data using t-statistic and neighbour joining algorithm. Dendogram using the t-statistic combined with the neighbour-joining algorithm, on the non-normalised cancer data. No general pattern is observed in terms of similar samples clustering together.

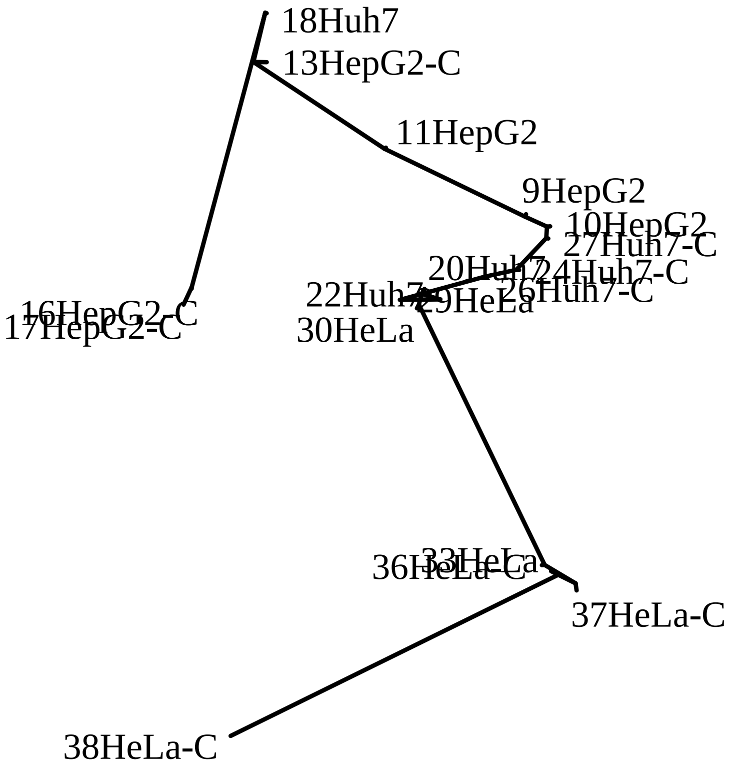


Figure 9.5. Grouping of normalised cancer data using t-statistic and neighbour joining algorithm. Dendogram using the t-statistic combined with the neighbour-joining algorithm, on the normalised cancer data. Good clustering is obtained for all groups (three types of cells, and the three sets of their controls.

9.4.2 k-means clustering analysis

k-means clustering analysis did not produce any interesting clustering results for the data, although when using three clusters, the normalised data, and a Manhattan distance metric, the HepG2 clustered together in a single group *along with the controls*, whereas the Huh7 and HeLa cells and Huh7 controls formed a second group, and the third group consisted entirely of HeLa controls. So in general one can say that k-means clustering analysis is not that useful on this data, and in the case of the particular result mentioned, the following conclusions can be drawn:

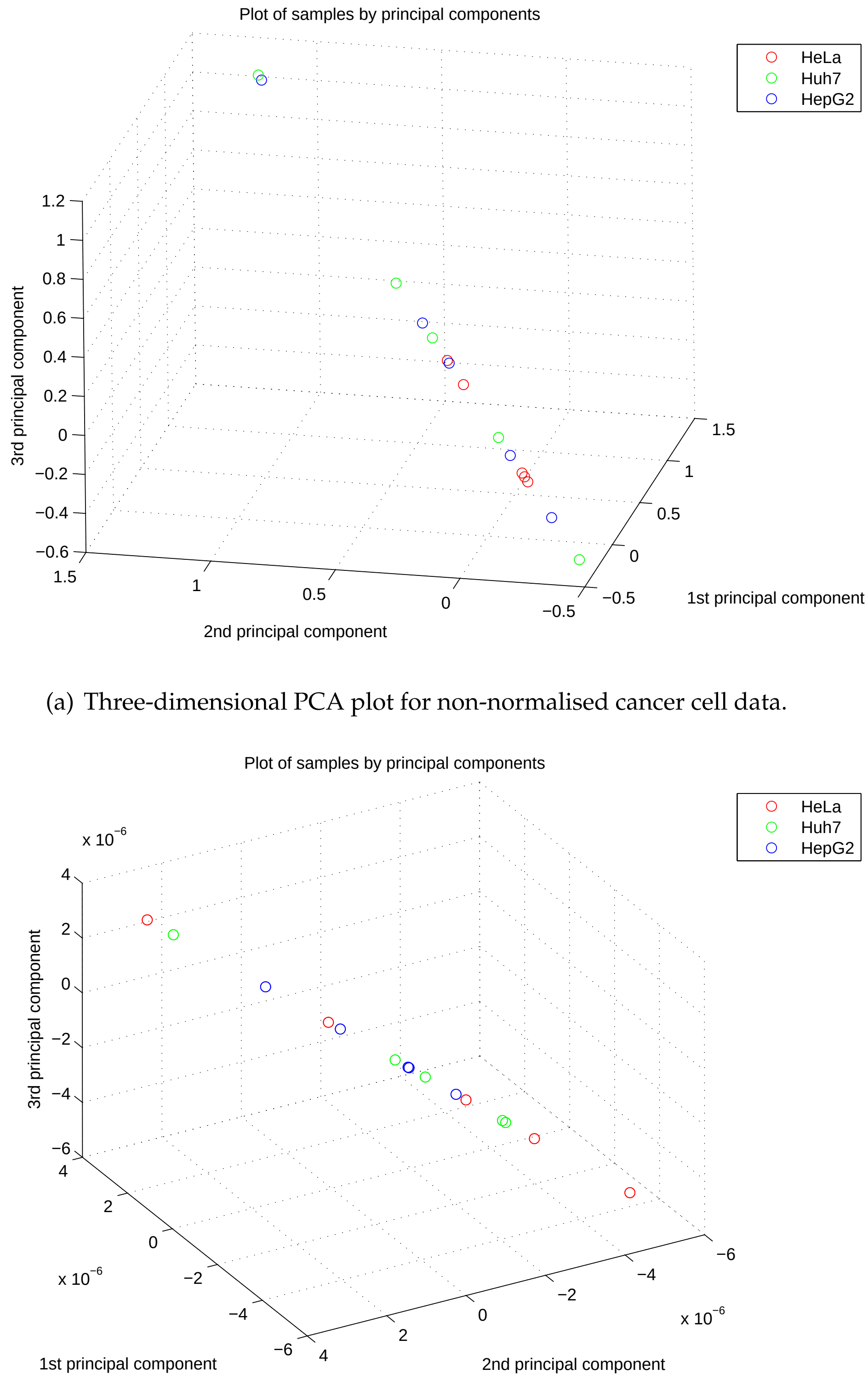
- The HepG2's show little variation from the controls, so perhaps the cells aren't producing many metabolites (that the mass spectroscopy can pick up).
- The HeLa cells are quite distinct from their controls.

9.4.3 Principle components analysis

Two-dimensional PCA plots failed to show any clear separation of samples in two dimensions. To ascertain if adding a dimensions helps separate the data, three-dimensional graphs were made (Figure 9.6). Again, neither graph shows anything like a clear separation of samples. Although higher-dimensional graphs cannot easily be made, it would seem unlikely that in any case they wouldn't show any clear separation of the data, except at very high dimensions, since the support vector machine can separate the data, although this may be using a nonlinear kernel which is difficult to visualise at higher dimensions.

9.4.4 Support vector machines

For the cell line media data, the $\boldsymbol{x}$ mentioned in the methods subsection on SVMs is the vector formed from the set of $I(x, t)$ mass spectroscopy intensity values for all mass/charge ratios x and all times t, and this is done consistently for all samples. As there were more than two categories, a variant on the simple two-class SVM mentioned in the methods must be used. The multi-class SVM method chosen uses a set of two-class SVMs, with labels $y \in \{0, 1\}$, to first distinguish whether a sample is in one group from both of the other classes (in this case of three classes) and then if it is in the



(a) Three-dimensional PCA plot for non-normalised cancer cell data.

(b) Three-dimensional PCA plot for normalised cancer cell data.

Figure 9.6. Three-dimensional PCA plots for the cancer cell data. No clear separation is shown in these 3D PCA plots between the three groups.

class of two classes combined, to decide which of those two classes it is in. To be robustly tested, the original set of 20x3 categories of samples were taken, and ten in each category were randomly selected for the training set and remaining ten were used for the test set. This was repeated ten times, using mutually unique training sets. For each repetition, 100% accuracy was obtained (ie no false positives and no false negatives), the support vector variable ρ is on the order of ± 5 (the sign being irrelevant), indicating a clear separation of the classes. If this seems a little too good to be true, then take it "with a grain of salt", since this is only very limited data (10 training samples + 10 test samples = 20 samples). However the experience with the autism study as detailed below shows that SVMs perform surprisingly well even when there is no clear visual separation of data in low-dimensional spaces.

9.5 Autism results

9.5.1 Pre-processing and preliminary analysis

The dendogram generated for the non-normalised spectra generally shows no significant grouping of autistic and non-autistic samples. Samples of autistic and non-autistic children were randomly scattered on the dendograms. However, it is interesting to observe that close relationships exist between the urine spectra of matched siblings. As revealed in Figure 9.7, samples A-24CM and N-24DM and samples A-70LH and N-70MH are both from two matched siblings respectively. It is noteworthy that the distance between the siblings of the same family is closer to each other compared to siblings from different families. Although consistency in this aspect is not observed in this study, the close relationships between the urine spectra of some matched siblings may suggest that urinary metabolic profile is strongly dependent on both genetic and environmental factors.

For the dendogram produced from the analysis of normalised spectra, close inspection of the data reveals some grouping of autistic samples into two separate clusters with a cluster of non-autistic samples also being separated out (Figure 9.8). Although there is some overlapping of sample points, it is clear that significant number of autistic and non-autistic samples does group up at certain areas on the dendograms. This observation may suggest that significant differences exists between the urine spectra of autistic and non-autistic children. Also, given that such separation is only visible after the spectra have been normalised, this may indicate that normalization of the data

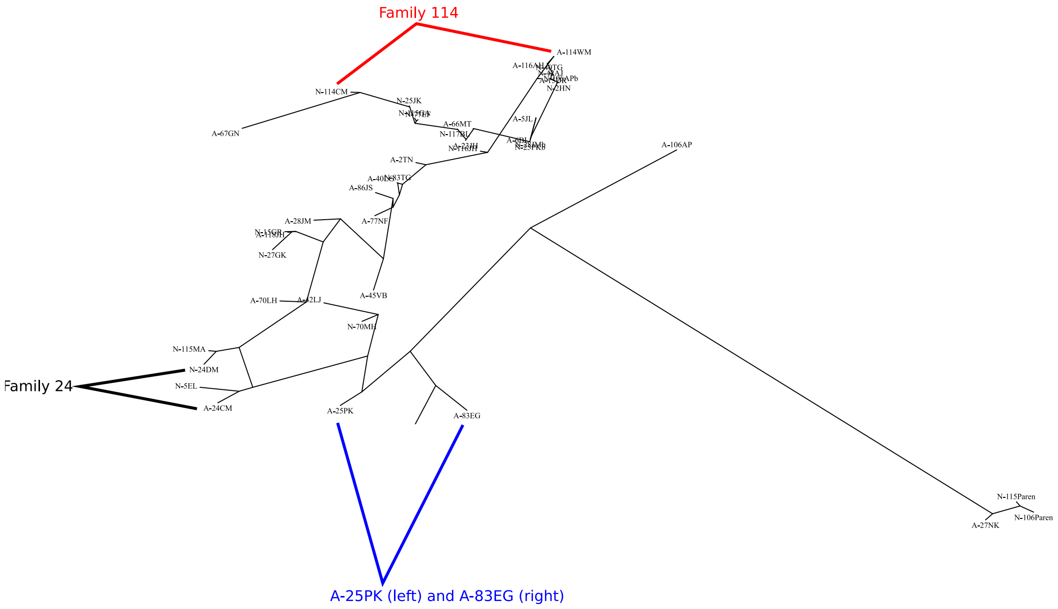


Figure 9.7. Dendogram of the non-normalised autism study data. Dendogram using the t-statistic combined with the neighbour-joining algorithm, on the non-normalised autism study data. The distance between samples is a measure of how distinct they are, as ascertained using the t-statistic.

prior to analysis is an important step. As shown in Figures 9.7 and 9.8, overlapping of sample points are prominent for both non-normalised and normalised spectra and this renders visual inspection of the data rather difficult. Given that the sample size of this study is relatively small and the fact that interpretation of data is already problematical owing to poor presentation of results using dendograms, researchers should err on the side of caution when using this method, should the current study be expanded to a larger population. However, this is not to say that the Students pair-t test and the neighbour-joining algorithm method did not provide us with useful data.

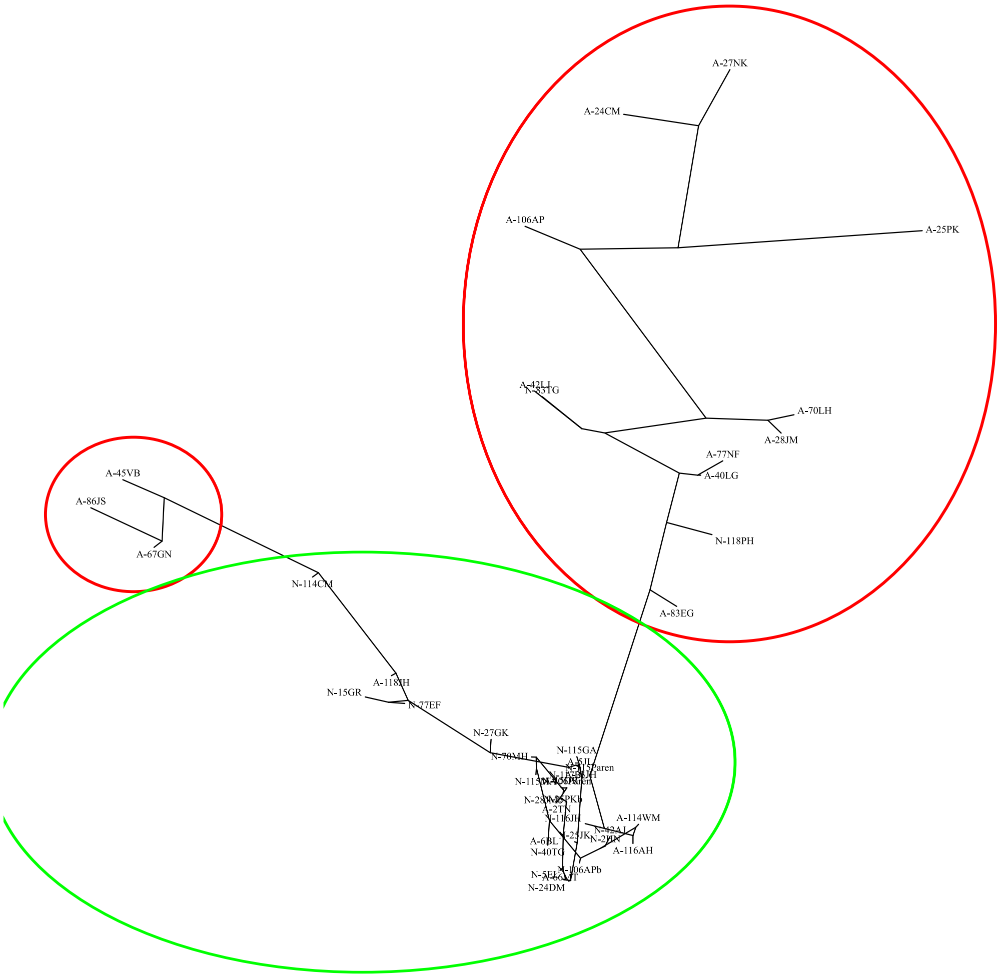


Figure 9.8. Dendogram of the normalised study data. Dendogram using the t-statistic combined with the neighbour-joining algorithm, on the normalised autism study data. The distance between samples is a measure of how distinct they are, as ascertained using the t-statistic.

Table 9.2. Sample classification using k-means clustering algorithm..

Category	Number of subjects	
	Cluster one (non-autistic)	Cluster two (autistic)
Non-normalised		
Autistic	16*	6
Non-autistic	13	9*
Normalised		
Autistic	17*	5
Non-autistic	17	5*

9.5.2 k-means clustering analysis

k-means clustering algorithm was used to group the urine samples of autistic children and non-autistic children into two clusters where cluster one and two were denoted as non-autistic and autistic populations, respectively. Findings revealed that there is no clear differentiation between the two clusters. In other words, it is impossible (at least for this data set) to differentiate between the non-autistic children from the autistic children based on this method (Table 9.2).

To put things into perspective, first consider the results yielded from the non-normalised spectra. Noting that cluster 1 shows a population of non-autistic children, 17 autistic children also fall into this category. There are 9 non-autistic children that fall into cluster 2, which denotes the population of autistic children. This is also true for the normalised data where 17 autistic children fall into the cluster of non-autistic children and 5 non-autistic children fall into those of autistic children. Ideally, the samples of autistic children and non-autistic children should be grouped into their respective cluster with minimal overlapping in each cluster. Therefore, the result does not clearly demonstrate to us that there is a distinct difference between the urinary metabolic profiles of both groups of children using k-means clustering algorithm.

9.5.3 Principal components analysis

Sample classification using the PCA is widely employed in the field of metabolomics study. The principal component (PC) scores plot is an efficient way to enable the visualization of any inherent pattern or clusters that may exist in the urine spectral profiles of autistic and non-autistic children. Any separation of data points can be attributed to differences between the urinary metabolic patterns of both groups.

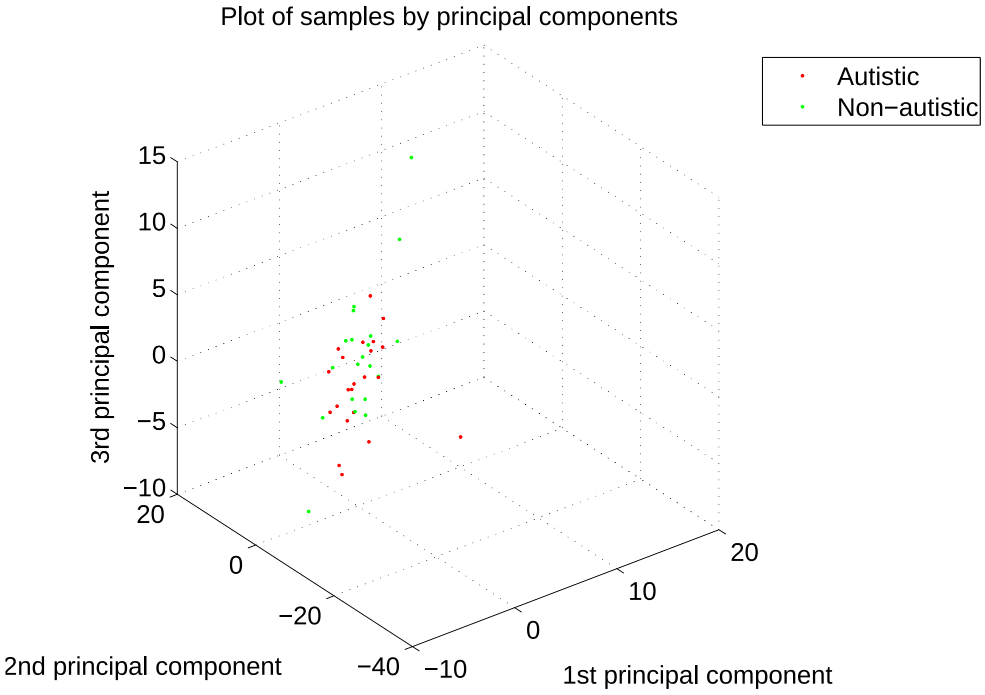
Two-dimensional PCA failed to demonstrate any significant segregation of data points into different clusters. Urines from autistic children did not reveal a urinary metabolic profile which is distinct from those of non-autistic children. In addition, the plots reveal that samples from the autistic group overlapped with samples from the non-autistic group. However, bear in mind that it will be difficult to determine with confidence that significant clusters, if any, do exists based on a 2D PCA. Any significant separation of data points which has not been revealed in the 2D PCA does not necessarily imply that there will not be significant clusters observed in 3 dimensions, as when plotted in space, data points may take different arrangements which will provide us with a clearer overview of the pattern that exists in the urine metabolic profile of autistic and non-autistic children. Hence, an additional third principal component was considered to generate a 3D PC scores plot for both non-normalised and normalized spectra as illustrated in Figure 9.9 In the case of this study, the resulting 3D PC scores plots did not show any clear separation either. This finding suggests that classification of samples into autistic and non-autistic groups have not been successful using the PCA.

9.5.4 Support vector machines

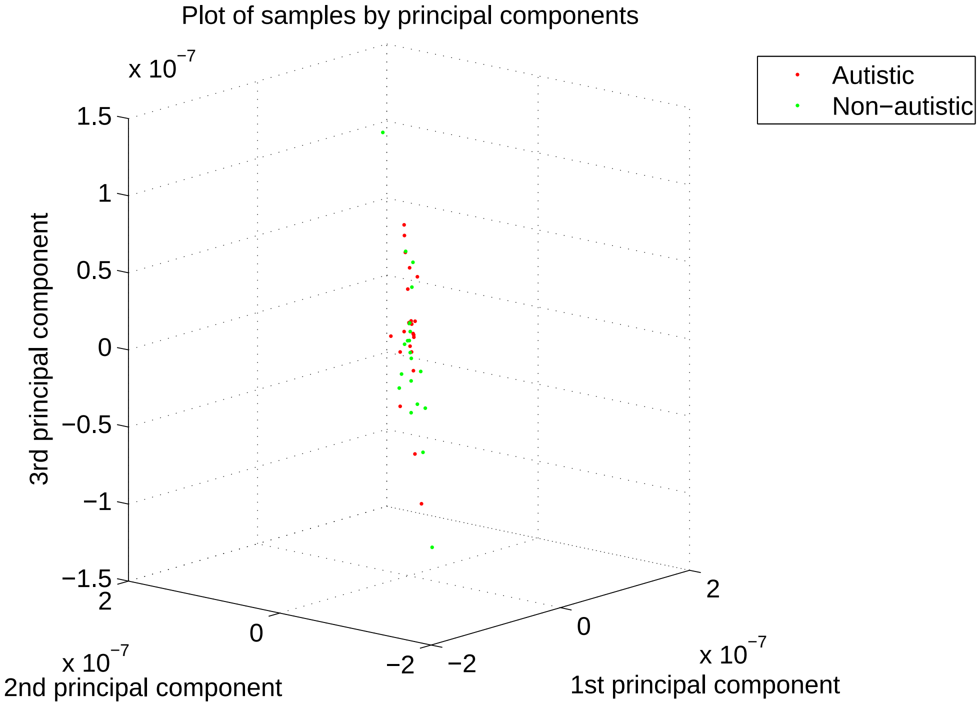
Although significant groupings were not observed with the above pattern recognition methods, a trial was performed using the SVM method. The SVM method was used in an attempt to distinguish between samples from autistic and non-autistic subjects, using the particular SVM methodology as detailed by Guermuer (2002). In this case, the data vectors were formed from the set of intensity values $I(x, t)$ for all mass/charge (m/z) values $\{x\}$ and all retention times $\{t\}$ and this was done consistently for all samples. The labels were denoted as

$$y = \begin{cases} 1, & \text{Autistic}, \\ -1, & \text{Non-autistic}. \end{cases} \tag{9.5}$$

To ensure the robustness of the learned model, 11 samples in each category were randomly selected for the training set and remaining samples (11 autistic and non-autistic samples respectively) for the test set. The machine learning was then performed 10



(a) Three-dimensional PCA plot for the non-normalised autism study data.



(b) Three-dimensional PCA plot for the normalised autism study data.

Figure 9.9. Three-dimensional PCA plots for the autism study data. Three-dimensional PCA plots for the autism study data. No clear separation is shown between the two groups.

Table 9.3. Success of SVM at distinguishing non-autistic from autistic children based on urine samples. This table shows the success rate at distinguishing autistic from non-autistic groups over 10 randomised choices of which 11 of the 22 autistic children are chosen, along with 11 of the 22 non-autistic children. The mean and standard deviation (SD) are also shown. These are very promising results.

Run	Non-normalised, linear kernel	normalised, polynomial kernel
1	90.91	95.45
2	95.45	68.18
3	81.82	90.91
4	86.36	68.18
5	90.91	77.27
6	86.36	86.36
7	90.91	77.27
8	95.45	68.18
9	90.91	59.09
10	100	90.91
Mean ± 1 SD %	90.91±5.25	78.18± 12.27

times using the same number of randomly selected training and test sets for each individual run. With each run of SVM learning using a set of randomly selected training samples, a learned prediction model was established. Ideally, this prediction model will be able to accurately assign an unknown sample to its true category, whether it is an autistic or non-autistic sample. The accuracy of this prediction model which was tested by the remaining samples yielded a mean success rate of 90.91 ± 5.25% (1 SD) using the non-normalised spectra and 78.18 ± 12.27% using the normalised spectra respectively. In the case of this study, prior to any SVM learning, training samples were first randomly selected from autistic and non-autistic groups to identify the best kernel type for both the non-normalised and normalised spectra. It was found that non-normalised spectra of autistic and non-autistic children were best separated using a linear kernel type as compared to a polynomial kernel when using the normalised spectra of both groups of children. Sample classification using SVM was subsequently performed for both non-normalised and normalised spectra based on the best kernel type that had been identified. The mean success rate was determined by taking the average of the individual success rates generated from 10 consecutive runs where each run was performed independently (Table 9.3).

The small SD of the success rate of the prediction model constructed using the non-normalised spectra suggests that this learned model is relatively more robust compared to that of the normalised spectra, the success rates of which tend to fluctuate between runs. Although the learned prediction model of the non-normalised spectra shows a high mean success rate and thus looks promising, this may be misleading as variations in dilution volume, which is a confounding factor, were not taken into account. Therefore, it is fair to say that the learned prediction model that was designed based on the normalised spectra is a better representative of the true value of SVM in classifying autistic and non-autistic children. Although the mean success rate of correct class prediction by SVM dropped after the spectra were normalised, the accuracy rate is still encouraging given that it is close to 80%.

9.6 Conclusions

With the exception of SVM, and in some cases the combination of the t-statistic with the neighbour-joining algorithm, the various pattern recognition methods used in this study failed to demonstrate clear separation of autistic and non-autistic children, and of the cancer cell data. With the combination of Students paired-t test and neighbour-joining algorithm approach, a close relationship between the urine metabolic profiles of some autistic and non-autistic matched siblings was prominent, and the non-normalised cell data was well-clustered. In the case of the autism data, although there is a lack of consistency in this observation across families, this result may indicate that urine metabolic profile is highly dictated by genetic and environmental factors. The k-means clustering algorithm and PCA were unsuccessful in differentiating urine samples of autistic from non-autistic children, and differentiating the cancer cell data. Both pattern recognition methods revealed a consistent overlap in the grouping of samples from the two populations. With the SVM approach, a high success rate of correct sample class prediction was achieved for both sets of data. The ability of SVM to correctly assign autistic and non-autistic samples despite the small number of samples used in this study indicates metabolomics is capable of discriminating autistic from non-autistic children based on the analysis of urinary metabolic profiles using SVM. A larger sample size in future studies may further enhance the accuracy rate of class prediction by SVM. As the cancer data shows, a large sample size dramatically improves the SVM success rate. Overall, this research shows that metabolomics has clinical utility in the diagnosis of autism, cancer, and other diseases.

To summarise the novel contributions of this work: algorithms and methods for processing (mass spectroscopy) metabolomic data were developed, some of these were used to ascertain if there are any statistical differences, and others were used in classifying samples from different sources.

Chapter 10

Conclusions

IN this thesis, analytical and modelling of complex systems for a large number of biological systems is developed. This chapter summarises the work, and draws some general conclusions.

10.1 Overview

This thesis tackles the following complex biological systems:

1. The evolution and analysis of DNA sequences.
2. The human brain during sleep.
3. Analysis of metabolites from various biological systems under different conditions (metabolomics).
4. Modelling of the development of tumors and cancer.
5. Transmission of viruses.

Topics 1, 4, and 5 all involved computer simulation and mathematical modelling of biological systems, exploring the overall behaviour of the system from the rules governing the interacting parts. Topics 1 through 3 involved statistical analysis of data from real biological systems. The following sections summarise the conclusions of each chapter.

10.2 Analysis of DNA sequences

In Chapter Two some existing methods for analysing DNA were reviewed, and showed some interesting microsatellite repeat patterns in *S. aureus*. The multifractal measure was combined with the minimal-span tree in a novel way to successfully classify bacteria. The Higuchi fractal measure (a complex systems tool) was extended to DNA sequences. Along with a measure of mutual information, this was applied to the analysis of correlations in DNA sequences generated in Chapter Three.

10.3 Mutations in DNA sequences

In Chapter Three the Higuchi fractal measure and mutual information measure of Chapter Two were then applied in order to study long-range correlations that can be found in short sequences of real DNA, "virtual" DNA, and throughout whole chromosomes. Genetic mutations were simulated for "junk" DNA sequences, with fill, copy, and mutate operations found to produce long range-correlations approaching 1024 bases in length. A negative test, with computer generated random sequences, succeeds in that no significant long-range correlations were found in these sequences. These results confirm that mutational events in non-conserved regions of DNA can give rise to long-range correlations.

10.4 Viruses and memes

In Chapter Four, two complex systems models were presented: one to capture the spread of viruses in a social network in a school, the other to capture the spread of memes in a social network of share market traders. Cycling patterns were found in both the model and the actual school data, strongly suggestive of successful capture of network structure and its impact on viral spread dynamics. In the share market model, it was similarly found that network structure impacts on the spread of memes through different groups of traders, and that this results in boom-bust cycles.

10.5 *Drosophila*

In Chapter Five, the focus was on modelling part of the gene network of *Drosophila* involved in setting up stripes in the larvae that regulate the future body plan. It was found that a cellular automaton is able to generate realistic patterns of stripes. This shows both the power of cellular automata in modelling complex systems, and in this specific case some of the robustness that has been evolved over countless generations of *Drosophila*.

10.6 *p53*

In Chapter Six, attention was paid to another gene network, that of the *p53* network in humans. *p53* is an important gene that (among other things) regulates cell repair and programmed cell death in cells. It was found a switching network based on a portion of the *p53* gene network allows for a good exploration of the choice between repair and cell death that cells face when encountering DNA damage. In related work looking at a model of mutations in *p53*, it was found that the number of inherited mutations in the *p53* gene plays a key role in early development of cancer.

10.7 Cancer

Chapter Seven then took a step up and looked at cancer as a whole, and explored facets of the multistep model of oncogenesis. The key findings of this work were:

1. The fastest path to somatic cancer is predicted to be through gaining mutations in D (evasion of cell death), then R (increased replication rate), then A (angiogenesis), then G (increased mutation rate).
2. Of the four categories of mutations, inheriting a mutation in G is predicted to produce cancer at the earliest age.
3. The fastest path to somatic cancer is robust to realistic changes in parameters, with the model that was developed being most affected by variations in tumour volume doubling time.

10.8 The human brain during sleep

In Chapter Eight, some procedures were established for cleaning data generated by a messy, biological, complex system: sleeping children. Some nonlinearity in (brain) processes generating EEG data was found, using both a time-reversal test and a fractal-based test. It was found, however that a significant portion of sleep EEG data can be considered to come from a linear process. This shows that both linear tools, such as the fast Fourier transform, and newer nonlinear time series analysis tools should be used. Due to the significant changes in nonlinearity between sleep stages, and the Higuchi fractal measure, one can be fairly certain that the nonlinearity process arises in the brain and not as a result of any nonlinear processes in the recording equipment.

10.9 Metabolomics

Chapter Nine detailed work in the field of metabolomics, looking (using mass spectroscopy) at the metabolic output of complex systems such as cancer and the human brain, and ascertaining useful data about these systems, even if it is unclear which of the metabolites is providing this information. Of all the tools tested, some (such as the t-statistic combined with the neighbour-joining algorithm) showed a little success at grouping samples from different sources, but none were as successful as the support vector machine (SVM) technique. SVMs had a high success rate of correct sample class prediction for both sets of data. The ability of SVM to correctly assign autistic and non-autistic samples despite the small sample size of this study indicates metabolomics is capable of discriminating autistic from non-autistic children based on the analysis of urinary metabolic profiles using SVM. A larger sample size in future studies may further enhance the accuracy rate of class prediction by SVM. Strong results were also found when applying SVMs to the problem of distinguishing cancer cell lines, based on (mass spectra of) samples of growth media in which they were growing.

10.10 Overall conclusions

It is not surprising that complex systems science enables a generic modelling and analysis paradigms for a large set of problems, and that these tools are highly useful; once one formalises the terms, it is really just applied mathematics. And mathematics is that most magical of tools that has unreasonable effectiveness in the natural sciences (Wigner 1960).

My thesis wouldn't be complete without quoting $\underline{\pi}$ (the movie) by Darren Aronofsky,

> *Restate my assumptions:*
>
> 1. *Mathematics is the language of nature.*
> 2. *Everything around us can be represented and understood through numbers.*
> 3. *If you graph the numbers of any system, patterns emerge. Therefore, there are patterns everywhere in nature.*
>
> *Evidence:*
>
> - *The cycling of disease epidemics.*
> - *The wax and wane of caribou populations.*
> - *Sun spot cycles.*
> - *The rise and fall of the Nile.*
>
> *So, what about the stock market? The universe of numbers that represents the global economy. Millions of hands at work, billions of minds. A vast network, screaming with life. An organism. A natural organism. My hypothesis: Within the stock market, there is a pattern as well... Right in front of me... hiding behind the numbers. Always has been.*