Industrial Mathematics Modeling Workshop for Graduate Students, July 22 - July 30, 2002

Edited by Pierre A. Gremaud, Zhilin Li, Ralph C. Smith and Hien T. Tran

Participants

Graduate Students

1. Ahmed, Suhail , Utah State University
2. Benedict, Brandy, North Carolina State University
3. Beun, Stacy, North Carolina State University
4. Brauss , Daniel , Clemson University
5. Buckingham , Robert , Duke University
6. Chaturvedi, Praveen, Southern Methodist University
7. Cho , Min Hyung , UNC Charlotte
8. David, John, North Carolina State University
9. Dogan , Gunay , University of Maryland
10. Enders , Joerg , Michigan State University
11. Feng , Jun , University of North Carolina, Chapel Hill
12. Gebauer , Bastian , Universität Mainz
13. Gishe , Jemal Emina , University of South Florida
14. Goncharov , Yevgeny , University of Illinois at Chicago
15. Hauck , Cory , University of Maryland
16. Jahan , Nusrat , Mississippi State University
17. Jin , Ya , Brown University
18. Kao , ChiuYen , University California, LA
19. Kas-Danouche , Said , NJ Institute of Technology
20. Kuster, Christopher, North Carolina State University
21. Latulippe , Joseph , Montana State University
22. Lee , Youngsuk , University of Wisconsin-Madison
23. Liao , Wenyuan , Mississippi State University
24. Lurati , Laura , Brown University
25. Martel , Rebecca , University of New Hampshire

26. Ouyang , Haojun , University of Toledo
27. Prodanovic , Masa , SUNY at Stony Brook
28. Rasmussen , Bryan , Georgia Tech
29. Robinson, Tracy, North Carolina State University
30. Silantyev , Valentin , Northeastern University
31. Stephenson , Rebekah , Michigan State University
32. Strain , Robert , Brown University
33. Tate , Brian , Texas Tech University
34. Vo , Tai Anh , California State University, Fullerton
35. Wang , Yuanyuan , University of Waterloo
36. Yuan , Shenglan , Auburn University
37. Yokley, Karen, North Carolina State University
38. Zhao , Ming , State University of New York, Stony Brook
39. Zhao, Yaxi, University of North Carolina Wilmington

Problem Presenters

1. Chen, Yu, SUM/MUS
2. Guilak, Farshid, Duke University Medical Center
3. Parham, Fred, NIEHS
4. Portier, Chris, NIEHS
5. Robertson, Lawrence (Robbie), US Air Force Research Lab, Kirtland Air Force Base
6. Royal, Tony, Jenike & Johanson
7. Whitaker, Shree, NIEHS
8. Williams, Pamela, Sandia National Labs
9. Young, Stan, CG Stat

Faculty Advisors

1. Banks, H.T
2. Gremaud, Pierre
3. Haider, Mansoor
4. Ito, Kazi
5. Li, Zhilin
6. Olufsen, Mette
7. Smith, Ralph
8. Tran, Hien

Contents

Preface

This volume contains the proceedings of the Industrial Mathematics Modeling Workshop for Graduate Students that was held at the Center for Research in Scientific Computation at North Carolina State University (NCSU), Raleigh, North Carolina, July 22 - July 30, 2002. This workshop which was the eigth one held at NCSU brought together 39 graduate students. These students represented a large number of graduate programs including

Auburn University, Brown University, California State University, Fullerton, Clemson University, Duke University, Georgia Tech, Michigan State University, Mississippi State University, Montana State University, NJ Institute of Technology, North Carolina State University, Northeastern University, Southern Methodist University, SUNY at Stony Brook, Texas Tech University, UCLA, UNC Chapel Hill, UNC Charlotte, UNC Wilmington, University of Illinois at Chicago, Universität Mainz, University of Maryland at College Park, University of New Hampshire, University of South Florida, University of Toledo, University of Waterloo, Utah State University.

The students were divided into seven teams to work on "industrial mathematics" problems presented by industrial scientists. These were not the neat, well-posed academic exercises typically found in coursework, but were challenging real world problems from industry or applied science. The problems, which were presented to the students on the first day of the workshop, required fresh insights for their formulation and solution. Each group spent the first eight days of the workshop investigating their project and then reported their findings in half-hour public seminars on the last day of the workshop.

The following is a list of the presenters and the projects they brought to the workshop.

- **Lawrence "Robbie" Robertson** (US Air Force Research Lab, Kirtland AFB) *Design of a membrane aperture deployable structure*
- **Pamela J. Williams** (Sandia National Laboratories) *Energy consumption and interference in the BART system*
- **Fred Parham, Chris Portier, Shree Whitaker** (NIEHS) *Mathematical Modeling of Comparative Initiation/Promotion Skin Paint Studies of B6C3F$_1$ Mice and Swiss CD-1 Mice.*
- **Farshi Guilak** (Orthopaedic Research Laboratories, Dept. of Surgery, Duke University Medical Center) *Mathematical models for articular cartilage: molecular diffusion in photobleaching experiments and signal transmission in a chondron*
- **Yu Chen** (Summus, Inc) *Recognizing sand ripple patterns from side-scan sonar images*
- **Tony Royal** (Jenike & Johanson, inc) *Surface profile of granular material around an obstacle*
- **S. Stanley Young** (CG Stat) *Predictive toxicology: benchmarking molecular descriptors and statistical methods*

These problems represent a broad spectrum of mathematical topics and applications. Although nine days is a short time for a full investigation of some of the aspects of such industrial problems, the reader will observe remarkable progress on all projects.

We, the organizers, strongly believe that this type of workshop provide very valuable non-academic research related experiences for graduate students while contributing to the research efforts of industrial participants. In addition, this type of activity facilitates the development of graduate students' ability to communicate and interact with scientists who are not traditional mathematicians but require and employ mathematical tools in

their work. By providing a unique experience of how Mathematics is applied outside Academia, the workshop has helped many students in deciding what kind of career they aspire to. In some cases in past workshops, this help has been in the form of direct hiring by the participating companies. By broadening the horizon beyond what is usually presented in graduate education, students interested in academic careers also find a renewed sense of excitement about Applied Mathematics.

The success of the workshop was greatly enhanced by active participation in a very friendly atmosphere and almost uninterrupted work during the nine days of attendance. The organizers are most grateful to participants for their contributions. The organizers would like to thank the National Science Foundation (Grant DMS-0204515), the Center for Research in Scientific Computation and the Department of Mathematics at North Carolina State University for their generous financial support. Special thanks are due to the faculty and staff of the Center for Research in Scientific Computation, the Department of Mathematics and North Carolina State University for the provision of excellent facilities and services. Finally, we would like to thank Brenda Currin, Kathleen McGowan and Rory Schnell for their efforts and help in all administrative matters. We are also grateful to Brian Lewis and Michael Zager for their help in providing transportation for the participants and to Terry Byron from the Department of Statistics for helping with setting up working computer accounts.

Pierre Gremaud, Zhilin Li, Ralph Smith, Hien Tran,
Raleigh, 2002.

Report 1

Design of a membrane aperture deployable structure

Joerg Enders[1], Said Kas-Danouche[2], Wenyuan Liao[3],
Bryan Rasmussen[4], Tai Anh Vo[5], and Karen Yokley[6]

Problem Presenter:
Lawrence "Robbie" Robertson
US Air Force Research Lab, Kirtland AFB

Faculty Consultant:
Ralph Smith
North Carolina State University

Abstract

Ultra-lightweight, membrane primary mirrors offer a promising future for space telescope technology. However, the advantages of the lightweight structure of the mirror are restricted by an extremely high susceptibility to microyield. Hence, careful packaging of the membranes is required when transporting mirrors of this type into space. Four packaging models, a cylindrical roll, an umbrella model, a multi-cut model and a single cut model, are presented and compared with each other. Factors such as curvature of the compressed membrane, stability after deployment, and the size of the launch vehicle are considered. All four packaging models appear to be feasible with certain materials and hence warrant physical testing.

1.1 Introduction and Motivation

As described in [2], there has been a dramatic improvement in technologies and concepts for large telescopes for both ground and space applications. However, the act of launching objects into space poses specific constraints on the structure and deployment of the cargo transported. Due to the high launch cost, ultra-lightweight, membrane primary mirrors have long been sought after by both NASA and the Department of Defense as a technology that could realize large aperture systems with low areal densities. Research on membrane structures

[1]Michigan State University
[2]New Jersey Institute of Technology and Universidad de Oriente, Venezuela
[3]Mississippi State University
[4]Georgia Institute of Technology
[5]California State University Fullerton
[6]North Carolina State University

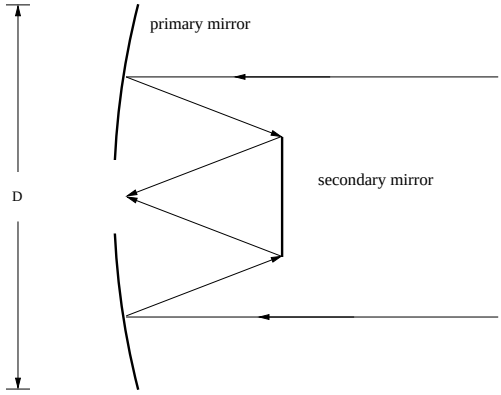


Figure 1.1: Schematic representation of membrane mirror system.

has culminated in the fabrication of meter-class lightweight structures with optical quality surfaces. These membranes are 10-100 microns thick and have surface qualities usable in visible spectrum applications, but the available structures that provide boundary support are so heavy that they eliminate the benefit derived from such lightweight apertures. Since these membranes must maintain an extremely high surface quality after release into space, the membranes cannot be packaged in ways that deform their shape outside an extremely small acceptable range. Thus, many factors must be taken into consideration when planning the folding of such membranes.

A sketch of a prototypical telescope used to resolve objects is provided in Figure 1.1. As the length of the diameter of the primary mirror increases, so does its power of resolution. Currently, the size of such telescopes has been bounded by the size of the rocket. More recently, however, researchers have begun to consider ideas regarding packaging methods that would enable the compactification of much larger mirrors without creating damage beyond desired accuracy. In order to attain the successful packaging of a large mirror, one must carefully consider the size of such an aperture, the size of the launch vehicle, the ease of deployment of the membrane into space, stability, the curvature of the folding method, as well as the allowable deformation of the material after being compacted.

Four different compacting schemes are considered in the following analysis. These schemes include folding arrangements for an uncut aperture as well as arrangements that require cutting the aperture at certain places. The radius of curvature necessary for each folding will be evaluated and compared to the minimum radius of curvature allowed by potential aperture materials.

1.2 Cylindrical Roll

As stated in the introduction, one goal of any membrane folding is to assure that it fits into the rocket. In this section and in Section 1.3, two ways to compress the membrane without cutting it are developed while noting that control of the maximal curvature is necessary.

The usable surface of the membrane also cannot be too small in order to guarantee a good resolution. The mathematical tool to compute whether the effective size of a membrane is sufficiently big is based on the *Modulation Transfer Function* (MTF). Consider the doubly curved membrane as depicted in Figure 1.2. Let D be the diameter of the aperture when projected into the $xy-$plane, and let d be the diameter of the hole in the membrane. If the matrix M has an entry of one wherever there is membrane material and zero elsewhere, the resulting *Autocorrelation Function* computes the convolution of M with itself. Dividing this matrix by the number of ones in M yields the MTF of M. If the MTF value at a point inside the perimeter of the original shape falls below 20%, it becomes difficult or even impossible to resolve certain objects. In Figure 1.3 values under 20% are black. Once those parts intrude into the black circle representing the size of the disk, the size of the surface is too small. In the case of the washer, d needs to be smaller than approximately $0.5D$ in order for the mirror to have a sufficient resolution. (Remark: In Figure 1.3 and later figures of the MTF, only the center part of the MTF matrix, which is the relevant part, is plotted.)

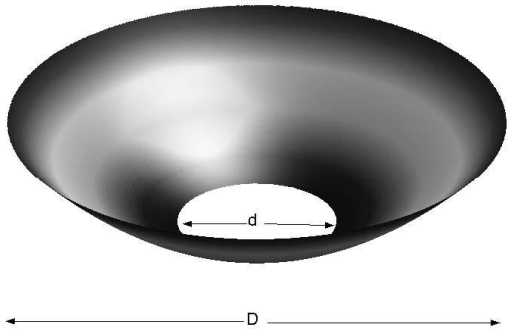


Figure 1.2: Doubly curved membrane.

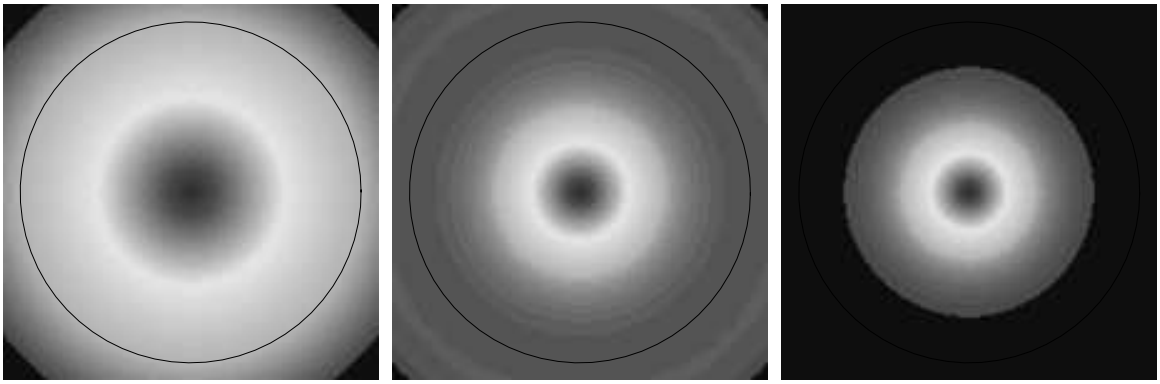


Figure 1.3: The MTF of the membrane matrix M for $d = 0.1D$, $d = 0.5D$, and $d = 0.6D$.

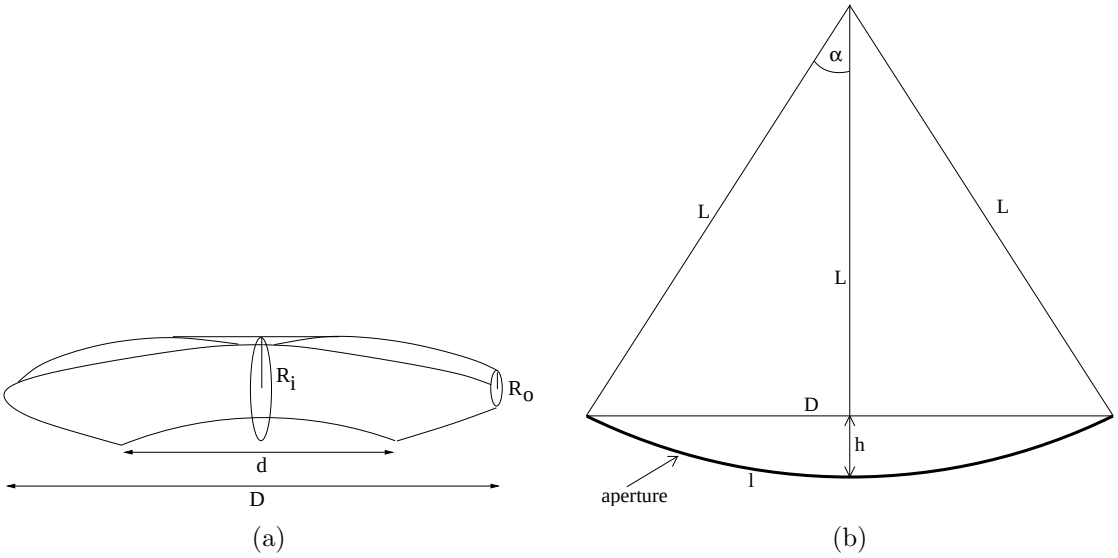


Figure 1.4: Cylindrical roll (a) and computing the different radii of curvature of the cylindrical roll (b).

Since the optimization of the total aperture weight is also desirable, the maximum d (i.e., $d \approx 0.5D$) is used in this and most of the other sections. However, one might want to go with a smaller d in order to increase stability.

When trying to find a way to fit the membrane into the rocket, probably the most straightforward idea is to roll the membrane once as drawn in Figure 1.4(a). If the membrane were flat, pulling two opposite sides of the aperture together would result in a cylindrical roll. In reality a more complicated shape would result because the mirror is doubly curved. Computation of the the maximal radius of curvature for this cylindrical roll begins by considering the radius R_i of the inner circle.

Assume that the aperture can be looked viewed as a circle, even if in fact it has a parabolic shape. This is reasonable as long as the radius of curvature L of the mirror is much bigger than its diameter D. From Figure 1.4(b) it follows that

$$\alpha = \sin^{-1} \frac{D}{2L}.$$

Therefore the arc length once across the membrane is

$$\ell = 2\alpha L$$

and hence

$$R_i = \frac{\alpha L}{\pi} = \frac{L}{\pi} \sin^{-1} \frac{D}{2L}. \tag{1.1}$$

From Figures 1.4 (a) and (b) one can see that when rolling the membrane, the outer radius of curvature R_o is bounded below by

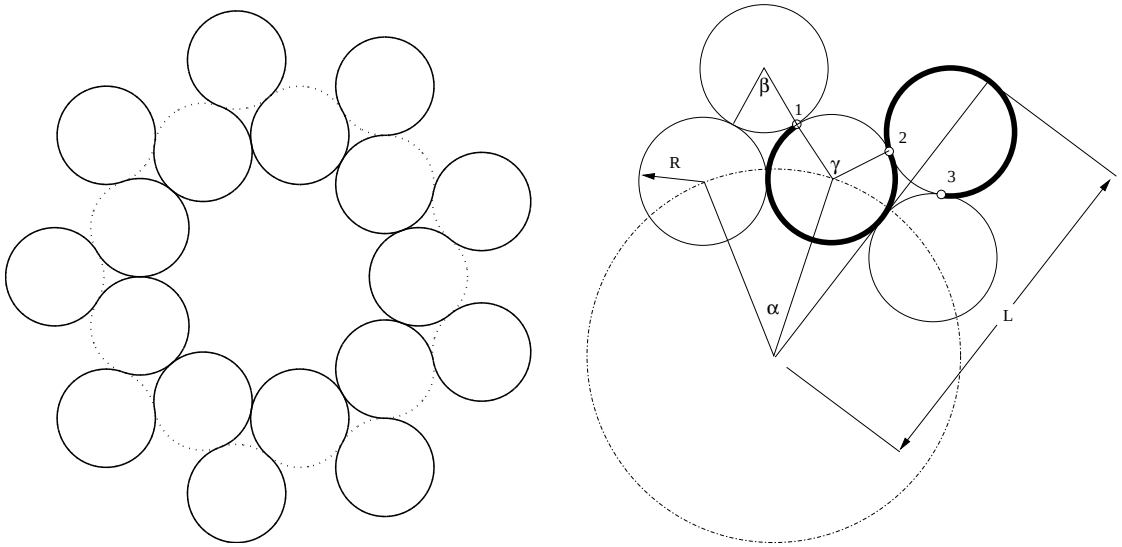
$$R_o \geq R_i - h$$

where

$$h = L - \sqrt{L^2 - \frac{D^2}{4}}.$$

Therefore the highest curvature will appear at the outside of the roll. Hence, for the minimal radius of curvature $R_f = \min\{R_o, R_i\} = R_o$ of the cylindrical roll, the estimate

$$R_F \geq \frac{L}{\pi} \sin^{-1} \frac{D}{2L} - L + \sqrt{L^2 - \frac{D^2}{4}} \tag{1.2}$$



(a) (b)

Figure 1.5: Cross-section of the folded umbrella (a) and analysis of the folded umbrella (b).

holds. Since the material can only sustain a limited curvature due to optical constraints, (1.2) sets bounds on L and D (See Section 6.4).

Finally, consider the constraint that the folded membrane has to fit into the rocket, and let R_{rocket} be the radius of the rocket. It follows from (1.1) that if the cylindrical roll is in an upright position, this condition is satisfied if

$$R_{rocket} \geq R_i = \frac{L}{\pi} \sin^{-1} \frac{D}{2L}. \tag{1.3}$$

Clearly the height of the roll is given by D.

If, for example, the rocket has radius $R_{rocket} = 2$ m and the mirror has radius of curvature of $L = 20$ m, then we can send a membrane mirror of 12.3 meters into space using the cylindrical roll.

1.3 Umbrella

The umbrella design, while one of the easiest to assemble, is also one of the least compact. Simply put, this design is a doubly-curved washer, folded down an axis through the center like an umbrella. The only action required to unfurl it is a one-dimensional slide along a rod. Like the cylindrical roll, the umbrella design requires no cuts of the material and can therefore tolerate a large inner diameter while maintaining a sufficiently dense modular transfer function.

Of course, there is no canonical way to fold a circular, doubly curved washer along the inner wall of a cylinder, so this design requires a choice of folding patterns. The following discussion is an analysis of one particular – and particularly straightforward – choice of folding geometry, but others may be more efficient.

Figure 1.5(a) shows a cross-section of the folded mirror with the proposed folding geometry. It is a set of nine circles arranged in a ring, with nine circles on the outside of the ring. As discussed below, this design may accommodate more or fewer circles, depending on the compactness requirements. The folded mirror weaves through the circles in a natural pattern, as indicated by the solid lines. Figure 1.6 shows a diagram of a full, three-dimensional, compressed mirror.

The use of circles in the folding pattern makes the overall shape easy to analyze. Figure 1.5(b) shows an expanded region of a cross section of the folded design. Let N be the total number of circles in either the inner or outer ring, and let D, d, and R_{rocket} be defined as in Section 1.2. A few trigonometric computations

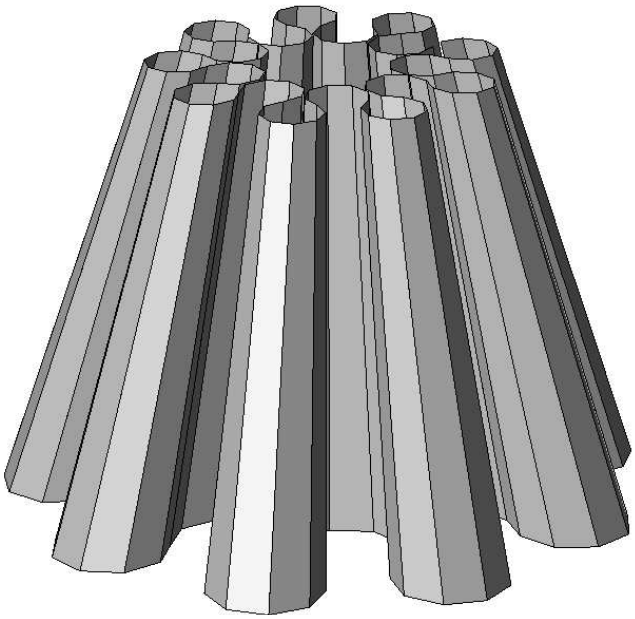


Figure 1.6: The folded umbrella

yield the angles and lengths

$$\alpha = \frac{2\pi}{N} \tag{1.4}$$

$$\beta = \frac{5\pi}{3} \tag{1.5}$$

$$\gamma = \pi\left(\frac{1}{3}+\frac{2}{N}\right) \tag{1.6}$$

$$L = R\left(1+\sqrt{3}+\cot\frac{\pi}{N}\right). \tag{1.7}$$

Thus, the arc length from point 1 to point 2 is $\pi R\,(5/3-2/N)$ and the arc length from point 2 to point 3 is $5\pi R/3$. The total arc length is

$$s = N\pi R\left(\frac{10}{3}-\frac{2}{N}\right). \tag{1.8}$$

The number of folds, N, should be as small as possible in order to keep the minimum radius of curvature, R_F, as large as possible.

Two geometrical constraints determine the minimum N. First, the total arc length at the bottom (i.e., the widest part) of the umbrella must be πD. Second, the folded design must fit inside a payload bay with radius, R_{rocket}, so $L \le R_{rocket}$, again at the bottom of the umbrella. The constraints, together with (1.7) and (1.8), show that N must be large enough to satisfy

$$R_{rocket} \ge \frac{D\left(1+\sqrt{3}+\cot(\pi/N)\right)}{(10N/3-2)}. \tag{1.9}$$

In order to support the membrane properly, N should be at least 5, but this does not usually constrain the design, since the size of the payload bay dictates that there be at least 8-9 folds in most cases.

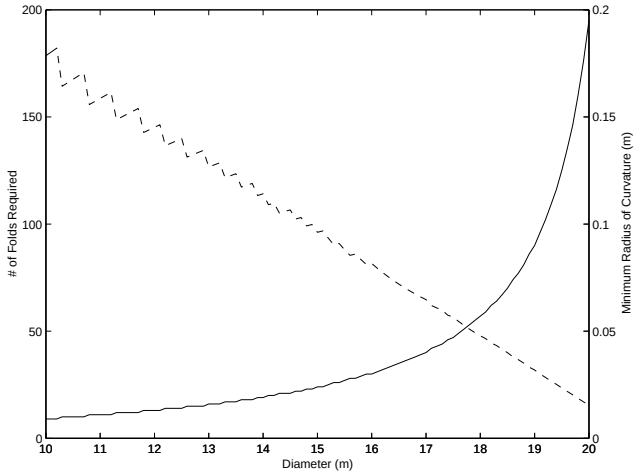


Figure 1.7: Design parameters versus D.

The total arc length on top is πd, so (1.8) gives the minimum radius of curvature

$$R_F = \frac{d}{(10N/3 - 2)}. \tag{1.10}$$

In summary, the design algorithm for the folding umbrella can be summarized as follows:

- Start with a given D and R_{rocket}. Choose d to be as large as possible while still satisfying the required optical properties (usually $d \approx 0.5D$).
- Use (1.9) to compute the minimum N required to fit the umbrella into the payload bay.
- Use (1.10) to compute the minimum radius of curvature at the top.
- This will dictate the material composition and maximum thickness of the membrane.

Figure 1.7 shows the behavior of N and R_F as functions of D for $d = 0.5D$ and $R_{rocket} = 2$ m.

Clearly, the number of folds becomes too large and the minimum radius becomes too small as D grows above 15 m. The reason for this is that the design uses only *circular* folds, whereas other choices might be more efficient in other situations.

Figure 1.8, for example, shows a cross section at the bottom of the umbrella for $R_{rocket} = 2$ and $D = 18$, which together force $N = 57$. The circular folds in this case are obviously inadequate, since the center of the cross section is an large open space with no material, while the folds form a highly twisted perimeter.

Therefore, for larger mirrors, other choices of folding patterns besides circles – such as stacking more layers of circles, using ellipses, etc., – are superior. As the folding patterns become more complicated, analysis of the design becomes more difficult.

1.3.1 Folding Pattern that Does not Work

To illustrate the possibility of other folding patterns, consider one choice that turns out to be inadequate. Let r_o, p, and θ be a fixed radius, a fixed integer, and an angular coordinate respectively. Define a cross-section of the umbrella as the image of a parameterized curve given in radial coordinates by

$$\mathbf{x}(\theta) = (r_0 + \rho \sin(p\theta), \theta). \tag{1.11}$$

The sine function applied along a circle becomes very sharp near the origin, thus inciting an unacceptably small radius of curvature. For example, consider $R_{rocket} = 2$ and $D = 10$, with $p = 10$. Then setting $r_0 = 1.25$

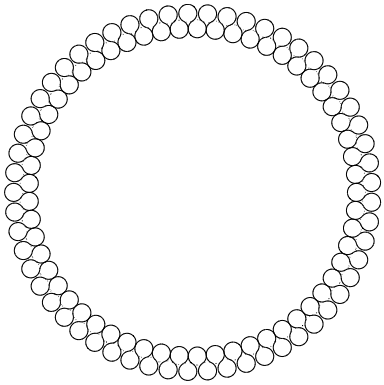


Figure 1.8: Failure of the circular folding pattern.

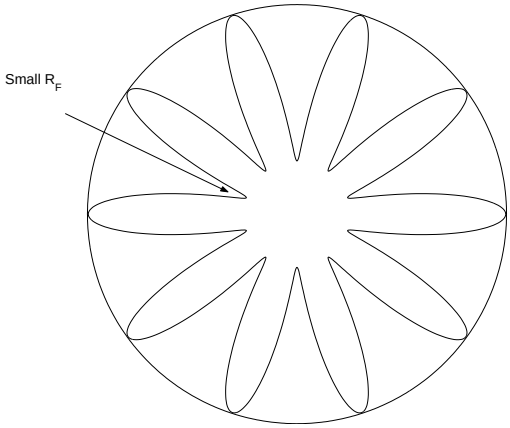


Figure 1.9: Parametric sine folding pattern.

at the bottom of the umbrella seems reasonable. The total arclength of the cross-sectional curve must be πD, which leads to a value of $\rho = 0.74$.

Figure 1.9 shows what happens next. A ring on the outside indicates the dimension of the payload bay, and the inner curve is the wrapping pattern. As the picture shows, the minimum radius of curvature that occurs nearest the origin is untenable. In fact, it has a closed-form representation:

$$R_{min,curve} = \min \left\{ \frac{(r_0 + \rho)^2}{r_0 + \rho + \rho p^2} \, , \, \frac{(r_0 - \rho)^2}{r_0 - \rho - \rho p^2} \right\}. \tag{1.12}$$

The minimum radius of curvature of the parametric curve with the given values is about 0.035 m – obviously too small for most materials.

1.4 Multi-cut Model

This section contains the analysis of a model which consists of dividing the membrane in N parts and then rolling them such that they can fit inside the rocket.

1.4.1 Analysis of one piece of the membrane

The multi-cut configuration is illustrated in Figure 1.10. Basically, N number of cuts are made along a diameter of the lens, from the outer disk to the inner one. This particular configuration allows an excellent original-to-

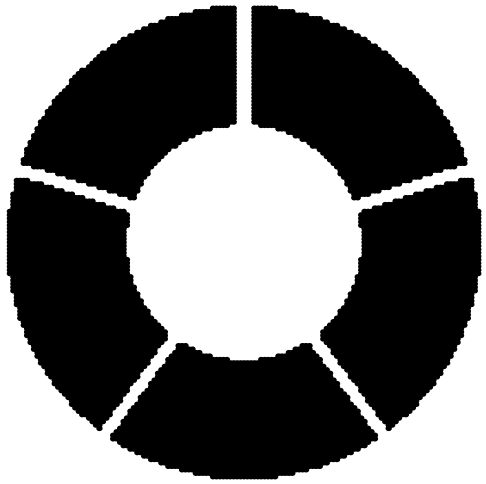


Figure 1.10: The multi-cut mirror with $N = 5$.

Table 1.1: Radius of curvature of each piece for different values of N (the number of pieces) and D (the aperture diameter)

N	Radius of curvature for $D = 10$	Radius of curvature for $D = 20$
2	1.59154943091895	3.18309886183791
3	1.37832223855448	2.75664447710896
4	1.12539539519638	2.25079079039277
5	0.93548928378864	1.87097856757728
6	0.79577471545948	1.59154943091895
7	0.69054741807754	1.38109483615507
8	0.60905959900277	1.21811919800554

packaged compression ratio, and thus larger lens would fit into current launch vehicles. The analysis of this configuration is not complicated and begins by reducing the three-dimensional lens down to a two-dimensional circle. This simplifying assumption is reasonable since the original three-dimensional lens curvature is low for telescope and focusing mirror applications.

As shown in Figure 1.10 and Figure 1.11, after a cut is made, the piece is rolled along chord S to form a shape that is circular near its bottom but somewhat parabolic at the top. The whole folded piece now looks like a cylinder with a diagonal part removed. The top portion probably would naturally go into a state of lowest energy, and the implication is that the curvature of the top is not high enough to cause concern. The bottom circle is where greatest curvature K_b will occur. This K_b can be expressed as

$$\begin{aligned} S &= 2\pi R_b &&= D\sin(\pi/N) \\ R_b &= \frac{D}{2\pi}\sin(\pi/N) \\ K_b &= \frac{1}{R_b} &&= \frac{2\pi}{D\sin(\pi/N)}, \end{aligned}$$

where N is the number of cuts, $N > 1$, and K_b is the curvature value at the bottom of the rolled up piece.

Since natural stability of the lens after deployment is a consideration, a large value of N would not be advantageous in that regard. However, cutting the lens into a large number of pieces does have its merit. When N goes past a certain value, pieces no longer need to be rolled for them to fit into the launch vehicle, and therefore curvature is no longer a consideration. For this particular configuration, a relatively small N in the range of three to eight was considered. Values of R_b for different N are summarized in Table 1.1.

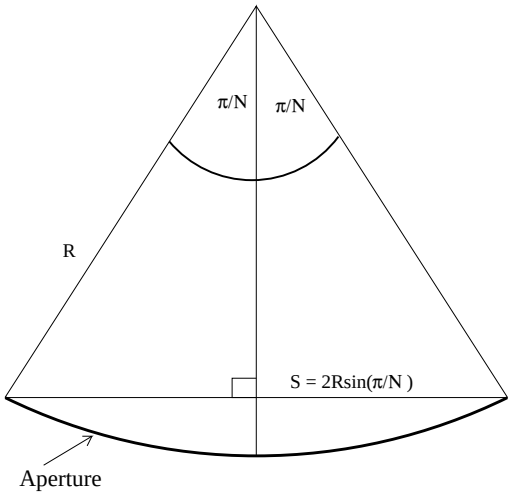


Figure 1.11: A single piece of the multi-cut mirror, viewed flat.

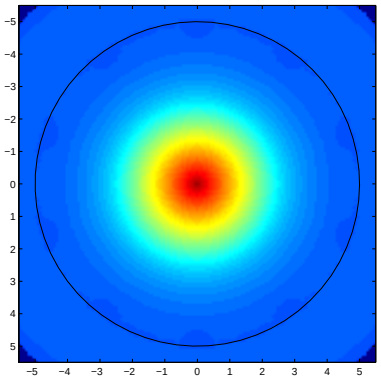


Figure 1.12: The MTF of the multi-cut mirror

Regarding the criterion concerning the *modulation transfer function* (MTF), Figure 1.12 shows that our lens achieves an overall value of 20% or better. Stability of the multi-cut model is questionable, but discussion of the natural frequencies of the mirror will not be presented here. To insure maximum stability for the multi-cut configuration, set the thickness, t of the material as high as possible. Curvature restraints dependent on thickness are contained in Section 6.4.

1.4.2 Packing and Deploying Procedure

The last subsection contained an analysis of each piece of the membrane, and the radius of curvature versus the maximum curvature of the material and the rocket radius can be used to decide the number of cuts to make in each membrane. Here, steps required in the packing and deployment procedure for the multi-cut model are summarized.

First the membrane is cut in N parts. Each part has a mechanism attached which acts as a sliding track. This will allow each membrane part to slide over the one next to it. This action is very similar to the mechanism used in sliding door closets. A string which connects and keeps all parts together goes through the holes on top of each sliding track. One part slides on top of the next one, then this set of two parts slides on top of the next part, and so on in successive process until all parts are in one stack. Figure 5.4 shows a stack of the membrane parts.

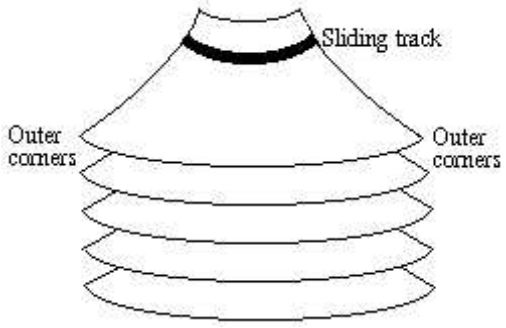


Figure 1.13: Stack of membrane parts.

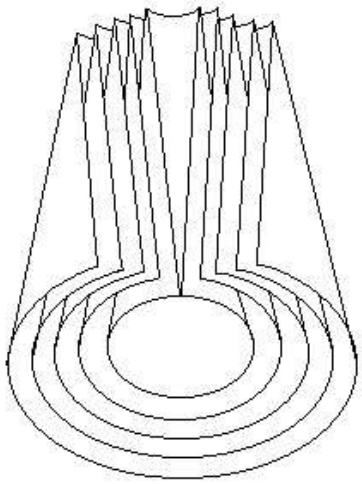


Figure 1.14: General view of all parts after the rolling process.

In the next step the membrane parts are rolled from top to bottom in such a way that the outer corners of the first part (on top) will touch each other, almost forming a cylinder. Next the other pieces roll around the first piece in a similar fashion.

A diagram of the whole set after the rolling process can be seen in Figure 1.14. A new concern appears in this process of packaging; the curvature of the most inner rolled part must satisfy the maximum curvature criteria, and the radius of curvature of the outermost rolled part must be less than the rocket radius.

In order to recover the original membrane, the unfolding proceeds in an inverse way: all the parts unroll and each part slides in the opposite direction as in the original packaging process. Finally, the string will tighten the separate pieces together.

1.5 The Single Cut Model

Cutting the disk in one place and then rolling the resulting strip around itself is another folding possibility. A sketch of the rolling model is contained in Figure 1.15. The single cut in the membrane can be described as a small width removed radially from the circle when evaluating the MTF, and the resulting information can be used to determine the largest usable diameter, d, for the hole in the center of the mirror. Using the largest possible d will help lower the radius of curvature of rolling the mirror around itself because the roll will likely have the highest curvature near the hole of the membrane. With only one cut, the mirror can retain more stability than in the multi-cut packaging scheme, but it obviously loses some of the stability of the original,

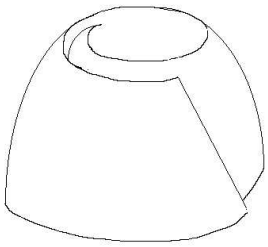


Figure 1.15: The single cut mirror rolled around itself.

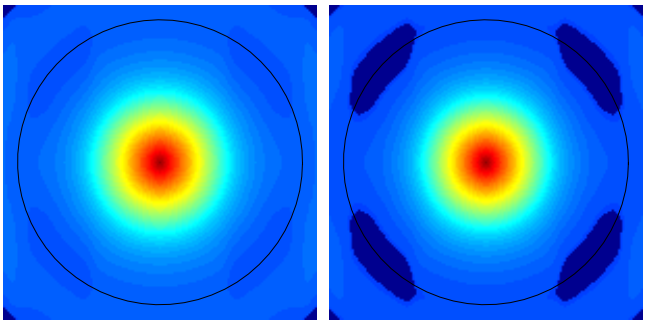


Figure 1.16: MTF of the aperture with $D = 10$ m and $d = 5$ m and $d = 5.5$ m, respectively.

uncut mirror in the umbrella and rolled folding patterns. Also, the single cut mirror rolled around itself would easily return to its original shape when released into space without the need of such elaborate support systems as in the multi-cut model.

A matrix describing the usable surface of the aperture can be generated and then sent to the MTF. A generous width of 0.3m was used to simulate the single radial cut for the MTF in order to estimate the maximum usable d.

The MTF of the matrix of an aperture with overall diameter 10 m and inner diameter 5 m is shown in Figure 1.16 on the left, and the MTF of an aperture with overall diameter 10 m and inner diameter 5.3 m is on the right. According to the MTF, a hole of diameter 5 m is acceptable, but black spots appear within the circle of the MTF on the right showing that a 5.3 m diameter for the inner circle is not feasible. A similar result is observed if $D = 20$ m and $d = 10$ m and $d = 10.7$ m, respectively, as shown in Figure 1.17. The maximum value of d for an aperture with a single cut appears to be close to $0.5D$ as in Section 1.2. Hence, the single cut does not seem to greatly affect the MTF.

Although the actual mirror would probably retain a parabolic shape when rolled, estimates for the curvature of rolling the mirror are quite easy to compute when considering the overall shape as a cone. The original curvature of the mirror is ignored when considering different lengths in the geometry of the cone model (see Figure 1.18). To insure that the cone fits within the rocket, R_b (the radius of the base of the cone) can be set equal to R_{rocket}. Further, R_t (the radius of the top of the cone) can be used as an estimate of the smallest radius of curvature in this packaging model, and hence, using similar triangles,

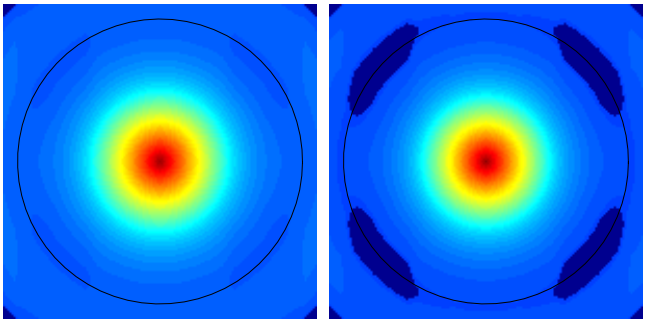


Figure 1.17: MTF of the aperture with $D = 20$ m and $d = 10$ m and $d = 10.7$ m, respectively.

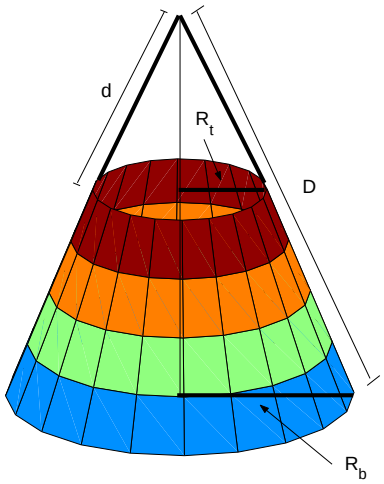


Figure 1.18: The geometry of the cone model.

$$\begin{aligned} \frac{R_t}{d} &= \frac{R_{rocket}}{D} \\ \Rightarrow R_F \approx R_t &= \frac{d}{D} R_{rocket}. \end{aligned}$$

Hence, the maximum curvature of rolling the mirror around itself can be estimated using the known constraint of rocket size, the diameter of the overall mirror, the diameter of the hole in the aperture found from the MTF, and simple geometry. Further, the inside of the rocket has known radius $R_{rocket} = 2$ m, and an estimate for R_t can be computed using the results from the MTF. Hence,

$$\begin{aligned} R_t &\approx 2 \cdot \frac{d}{D} \\ &\approx 2 \cdot \frac{D/2}{D} = 1, \end{aligned}$$

and further, $R_F \approx 1$.

One could consider other single cut packaging methods, such as using a parabolic function instead of a cone or stretching the mirror out into more of a spiral. The parabolic idea does warrant more study, but a spiral wrapping scheme may require too high a curvature for any accessible materials.

Table 1.2: Minimum allowable radii of curvature for different materials and different widths for the cylindrical roll, single cut, and multi-cut models.

thickness (t)	R_F for 2014-T6 Aluminum	R_F for I-400 Beryllium	R_F for 304 Stainless Steel
$10\mu m$	0.0012 m	557.3210 m	0.1187 m
$20\mu m$	0.0026 m	3340.2 m	0.2990 m
$30\mu m$	0.0040 m	9520.7 m	0.5135 m
$40\mu m$	0.0055 m	20018 m	0.7535 m
$50\mu m$	0.0069 m	35627 m	1.0147 m

1.6 Results and Conclusions

The estimated R_F values for various materials were computed using a formula from [1],

$$w \approx \left(\frac{t}{2R_F}\frac{E}{H}\right)^{\frac{1}{n}} \frac{\hat{D}^2}{4t},$$

where t is the thickness of the aperture, R_F is the radius of curvature of the folding model, E is the elastic modulus of the material, n is the strain hardening exponent, H is the plasticity model constant, $\hat{D}$ is the length of the surface being curved, and w is a measure of deflection. The value of w should be kept very low because the mirror will not reflect properly after even small deformations. As a result, $w = 1\mu m$ was deemed an appropriate estimate of allowed deflection and was used to approximate the minimum R_F value (or alternatively, the maximum curvature, $1/R_F$) that the specific material can be shaped to hold without losing the necessary properties of the mirror. For the single cut and multi-cut models as well as the cylindrical roll model, $\hat{D}$ can be estimated by the circumference of the circle with highest curvature. Hence, in these models, $\hat{D} = 2\pi R_F$. The equation from [1] can then be rewritten

$$\begin{aligned} w &\approx \left(\frac{t}{2R_F}\frac{E}{H}\right)^{\frac{1}{n}} \frac{(2\pi R_F)^2}{4t} \\ \left(\frac{4tw}{4\pi^2 R_F^2}\right)^n &\approx \frac{t}{2R_F}\frac{E}{H} \\ \left(\frac{(tw)^n}{\pi^{2n}}\right)\frac{2H}{Et} &\approx R_F^{2n-1} \end{aligned}$$

and therefore the minimum allowable radius of curvature can be computed using thickness and material properties. Using the outside loop of the umbrella base curve model, the value $D = \frac{5}{3}\pi R_F$ can be used to approximate the minimum allowed R_F for different materials by using

$$\begin{aligned} w &\approx \left(\frac{t}{2R_F}\frac{E}{H}\right)^{\frac{1}{n}} \frac{(\frac{5}{3}\pi R_F)^2}{4t} \\ \left(\frac{36tw}{25\pi^2 R_F^2}\right)^n &\approx \frac{t}{2R_F}\frac{E}{H} \\ \left(\frac{36tw}{25\pi^2}\right)^n \frac{2H}{Et} &\approx R_F^{2n-1}. \end{aligned}$$

Using Equations 1.13 and 1.13 and material constants from [1], the minimum allowable radii of curvature for the single cut, multi-cut, and cylindrical roll models for different materials were computed and are summarized in Table 1.2 for various thicknesses of the aperture.

Estimations for the minimum allowable radii of curvature for the umbrella model are contained in Table 1.3. These values can be compared to the radii of curvature needed for each model to fit within the rocket to decide

Table 1.3: Minimum allowable radii of curvature for different materials and different widths for the cylindrical roll, single cut, and multi-cut models.

thickness (t)	R_F for 2014-T6 Aluminum	R_F for I-400 Beryllium	R_F for 304 Stainless Steel
$10\mu m$	0.0012 m	312.8706 m	0.1051 m
$20\mu m$	0.0025 m	1875.1 m	0.2648 m
$30\mu m$	0.0039 m	5344.8 m	0.4547 m
$40\mu m$	0.0053 m	11238 m	0.6673 m
$50\mu m$	0.0068 m	20000 m	0.8985 m

if each scheme is usable. As long as a particular R_F from the tables is smaller than the R_F needed for a model, that material at that thickness will work for the model being considered. For all apertures that are 10-50 μm thick, both aluminum and stainless steel appear to be feasible materials if using any of the presented models with appropriate choices for the number of cuts in the multi-cut model and for the number of folds in the umbrella model. Beryllium does not appear to be an appropriate material for the aperture, but the physical properties of beryllium make it a less than favorable choice for the mirror regardless of curvature. Adequate stability of the aperture can possibly be achieved simply by maximizing the thickness of the mirror within the constraint that the radius of curvature of the model is larger than the radius of curvature allowed by the material.

Since the models are simplifications, physical testing is definitely suggested. Despite some of the simplifications with regard to original mirror curvature, all models presented do warrant the further study of physical models. The value for allowable deflection of the mirror, w, should also be thoroughly tested and may vary with material. Higher stability is achieved by higher natural frequency, and further study with these models with regard to natural frequency is suggested as well. Also, more research on possible support structures for the mirror would help the analysis of optimal mirror packaging.

Acknowledgments

The authors would like to thank the NCSU Center for Research and Scientific Computation, the NCSU Department of Mathematics, SAMSI, NSF, and everyone who helped with the 2002 Industrial Mathematics Modeling Workshop. The authors would also like to thank Robbie Robertson and Kirtland AFB for presenting such an interesting problem.

Bibliography

[1] J.L. Domber, L.D. Peterson, *Implications of Material Microyield for Gossamer Optical Reflectors,* 43rd Structures, Structural Dynamics and Materials conference, Denver, AIAA 2002-1503, April 2002.

[2] M. S. Lake, L. D. Peterson, M. B. Levine *A Rationale for Defining Structural Requirements for Large Space Telescopes,* 42nd Structures, Structural Dynamics and Materials conference, Seattle, AIAA 2001-1685, April 2001.

Report 2

Energy consumption and interference in the BART system

Brandy Benedict[1], Min Hyung Cho[2], Jemal Emina Gishe[3],
Rebecca Martel[4], Robert Strain[5], Brian Tate[6]

Problem Presenter:
Pamela J. Williams
Sandia National Laboratories

Faculty Consultant:
Mette Olufsen

Abstract

Train passengers expect a smooth, comfortable ride and timely arrival to their destination. However, commuters on the Bay Area Rapid Transit system often experience jarring rides as trains speed up and slow down to avoid colliding with the train in front of them and still stay on schedule. This interference between the trains can result in customer dissatisfaction as well as increased fuel costs. We develope a model to reduce interference by minimizing the energy consumed during acceleration. Using nonlinear optimization methods, we obtain results that show less interference when the variation in acceleration over a short time interval is limited. We also observe the effects of changing the length of the timestep, the initial conditions, and the length of a train on the amount of interference between two trains.

2.1 Introduction

Commuters in the San Francisco Bay Area expect a dependable train system; one which is capable of transporting them in a timely and comfortable manner. The automatic train control system for Bay Area Rapid Transit (BART) has the important job of controlling the train in a fashion that will transport passengers to their destinations as quickly and smoothly as possible. A passenger might assume that the train will travel as fast as possible in order to arrive at each station at the scheduled time. However, safety regulations require that a minimum distance be maintained between any two trains traveling in the system, sometimes a hard

[1]North Carolina State University
[2]University of North Carolina at Charlotte
[3]University of South Florida
[4]University of New Hampshire
[5]Brown University
[6]Texas Tech University

distance to maintain when station spacing is close, in areas like downtown San Francisco. Thus, when a train approaches too closely to the train ahead of it the control sytem commands the train to decelerate in order to maintain a safe distance from the lead train. This situation occurs frequently near stations, where the lead train makes its scheduled station stop for passenger boarding and trains behind it are drawing nearer to the station. When the lead train accelerates as it pulls out of the station, the following train will also accelerate, as long as the safe distance is maintained. There are two difficulties associated with this situation: the calculation of the required safe following distance, and the levels of acceleration to maintain that distance.

The current automatic train control system (ATC) for San Francisco's BART district uses hard-wired circuitry to determine train locations and communicate with trains. The ATC has limited capabilities for locating trains and controlling speed. There are only 8 selectable speed commands, which results in trains often traveling below the safety-enforced speed limits. Additionally, the current control system, as well as any future system, computes commands for a train's acceleration by assuming that the train immediately ahead of it is stationary. This situation results in frequent switching between acceleration and deceleration for the following train, a phenomenon termed *interference*. This cycle of deceleration and acceleration may continue for the full length of the line, increasing energy consumption by the engine as well as passenger irritation.

A new generation of train control systems called communication based train control (CBTC) is under development. Features of the CBTC system include radio communication between trains, more accurate train location and the ability to use increment speed commands every 1 mph. A system that allows for finer control commands reduces the necessary distance between two consecutive trains. This, in turn, reduces interference since trains that are sufficently far apart they do not affect each other's behavior. Reductions in speed limits also allows for reduced headway and limits interference, but reducing speeds adds to trip time, which is unacceptable both to the BART district and passengers. The train control system we consider here is within the context of the new CBTC.

In this project we will minimize the energy consumed. By doing so, we can cut fuel costs as well as increase customer satisfaction through a smoother ride and adherance to the established train schedule.

2.2 Problem Statement

Does minimizing the energy consumed during acceleration minimize the interference? To answer this question, we developed a simplified model of the BART system. Considering only two trains traveling through one station, we formulate an energy cost function to be minimized, subject to appropriate motion and boundary constraints. Using optimization tools, we seek to minimize the cost function; i.e. minimize the function describing work done by the train, and compare the resulting graphs of acceleration and velocity over time to determine the success of our model.

There are several ways that one may define interference. One posibility would be a limit on the number of jumps between acceleration and deceleration on a given section of track. However, this option would exclude the possibility of allowing a large number of very small jumps, which go largely unnoticed by passengers. We want to allow for this possibility. An unacceptably large degree of interference consists of jumping from a large acceleration rate to a large deceleration rate in some short period of time. In this formulation the number of such jumps is irrelevant. Thus, we define interference as the amplitude of the oscillations in acceleration that occur over a period of time. If our model produces a level of interference with a number of jumps of smaller amplitude, we consider this to be an improvement. Within this framework, there are several aspects we want to address:

1. How does energy consumption vary when the rate of change of acceleration is constrained, thereby forcibly reducing interference?

2. How does variation in train length affect interference levels?

3. How does varying the speed limit affect the level of interference?

4. How does varying the train schedule affect interference?

2.3 Methods

2.3.1 Assumptions

Energy is consumed as a train accelerates. We therefore define energy consumption as the amount of work done by the train during acceleration and velocity maintenance. We neglect deceleration, as energy is not used by the motors during deceleration. By minimizing the energy consumed we hope to obtain the optimal level of overall acceleration/deceleration and velocity necessary to maintain the safe distance between trains, thus reducing interference. This will result in significant financial savings for BART and also provide a more comfortable ride for passengers. We have previously generated data containing the position, velocity, and acceleration of the first train and are attempting to minimize the energy the second train uses in response to the actions of the first train. We assume the following in our model.

- The system consists of two trains and one station.
- Train tracks are straight and on level terrain.
- Tracks are covered; weather effects, road crossings, and debris can be neglected.
- Effects of friction and air resistance are neglected.
- Trains have identical attributes (length, mass, and engine capabilities).
- Trains are capable of attaining any acceleration value in a continuous range.
- Station location is fixed.
- Position, velocity, and acceleration of the first train are given.
- A train's wait time at the station (referred to as dwell time) is constant.
- Trains always adhere to the schedule; there are no delays.
- The train controller has perfect knowledge of the system: the exact location, velocity, and acceleration of all trains are known at all times.

To begin working towards our ultimate goal of minimizing interference on the BART system, we considered a simplified situation. Adding complexity to the system later will be a relatively easier task. For our problem to be computationally feasible using the Matlab optimization toolbox, we could only consider a track with one station and two identical trains, in which we knew exactly what the trains were doing at all times. With proper choice of initial conditions and other variables, interference is still observed near the station in the unconstrained model. In this initial formulation of the train system, we did not include any factors that could contribute to delays. As anyone who has used public transportation knows, delays in scheduled arrival time are not infrequent. A more accurate model would be probabilistic rather than deterministic, and include an element of randomness in travel time. This variation could be due to deceleration at road crossings or for debris on the tracks, or fluctuations in engine power. However, we believe the model we have created is a reasonable one, and will provide some insight on the reduction of train interference.

2.3.2 Mathematical Formulation

Our goal is to minimize the total energy consumed by the second train within the system. We begin with the trajectory information for the first train known completely. We define energy consumption as the work done by a train during acceleration and velocity maintenance [3], since no energy is consumed during deceleration. We define the positive part of acceleration for train 2 by

$$a_2^+(t) = \frac{1}{2}\Big[a_2(t) + |\, a_2(t)\,|\,\Big],$$

where $a_2(t)$ is the acceleration of train 2 at time t. Before stating the problem, we define the variables and parameters necessary for our formulation:

$x_i(t)$: Position of train i at time t [ft]
$v_i(t)$: Velocity of train i at time t [ft/s]
$a_i(t)$: Acceleration of train i at time t [ft/s^2]
T_2 : Time for train 2 to travel from the start point to the station
v_{max} : Speed limit
L : Distance from start point to station
h : Time step size for discretized numerical computation
ε : A parameter defining the allowed tolerance for acceleration
trainlength : Length of train

To minimize the work integral defined [3] by

$$\int_0^{T_2} a_2^+(t)v_2(t)dt, \tag{2.1}$$

subject to the constraints:

$$\begin{aligned}
\dot{x}_2(t) &= v_2(t) && (2.2)\\
\dot{v}_2(t) &= a_2(t) && (2.3)\\
x_2(0) &= 0 && (2.4)\\
v_2(0) &= 0 && (2.5)\\
a_2(0) &= 2.2 && (2.6)\\
x_2(T_2) &= L && (2.7)\\
v_2(T_2) &= 0 && (2.8)\\
a_2(T_2) &= 0 && (2.9)\\
0 \le v_2(t) &\le v_{max} && (2.10)\\
-3.2 \le a_2(t) &\le 4.4 && (2.11)\\
x_1(t) - x_2(t) &\ge trainlength + 50 + \frac{v_2^2(t)}{9} + 8v_2(t) && (2.12)\\
|a_2(t) - a_2(t+h)| &\le \varepsilon h && (2.13)
\end{aligned}$$

Equations (2.2) and (2.3) are Newton's equations of motion, governing the movement of the trains along the track and equations (2.4)-(2.9) are boundary conditions for the three main variables. Equations (2.10) and (2.11) are upper and lower bounds on velocity and acceleration respectively, which are determined by the maximum speed limit on a particular length of track, and the propulsion and brake limitations of the trains. Inequality (2.12) is the minimum safety distance between the front ends of two trains, which includes the length of the train, the 50 feet train buffer distance, and the braking distance of the second train dependent upon its velocity. Inequality (2.13) limits the rate of change of acceleration, to within a small tolerance ε .

For our simulation, we used the following parameter values:

$$\begin{aligned}
T_2 &= 60\\
v_{max} &= 88\\
L &= 2640\\
h &= 0.5\\
trainlength &= 700
\end{aligned}$$

Our implementation assumes that train 1 has already left the starting point and has been sitting at the station for 15 seconds before train 2 begins to accelerate. When train 2 begins its trip, train 1 dwells for

another 15 seconds in the station before continuing its journey. It may be assumed that train 1 is continuing along the track to its next station stop, but this detail is beyond the scope of our simulation. We also require that train 2 reach the station 60 seconds after the start of the simulation. Our paper focuses on the effects of varying the parameter ε on the energy consumption of train 2.

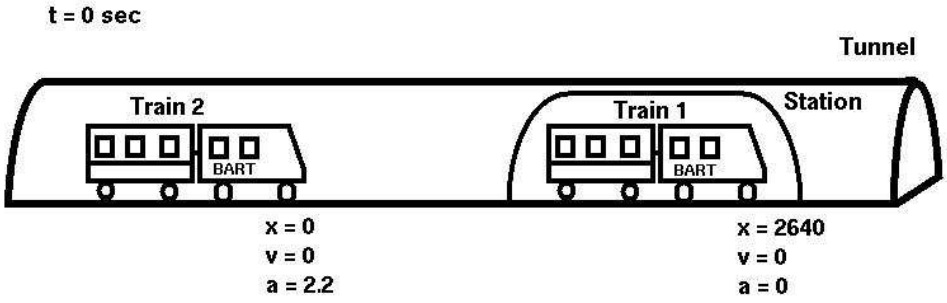


Figure 2.1: Position of trains at time zero

2.3.3 Numerical Method

In order to solve this optimization problem, we discretized both the objective function (2.1) and the equations of motion defined in the constraints. We use the trapezoid rule to discretize the integral in the objective function, giving:

$$\int_0^{T_2} a_2^+(t)v_2(t)dt \approx \frac{T_2}{2M}\left[a_2^+(t_0)v_2(t_0) + 2\sum_{j=1}^{M-1} a_2^+(t_j)v_2(t_j) + a_2^+(t_M)v_2(t_M)\right],$$

where $t_j = jT_2/M$, where M is the number of grid points. To implement this discretization, we employ the Matlab function `trapz`, which determines the approximate integral given the velocity and acceleration at each time step. To discretize the equations of motion, we use the backward Euler method, yielding

$$\begin{aligned} t_j &= t_{j-1} + h \\ x_2(t_j) - x_2(t_{j-1}) - v_2(t_j)h &= 0 \\ v_2(t_j) - v_2(t_{j-1}) - a_2(t_j)h &= 0 \end{aligned}$$

This constraint is implemented as a matrix equation of the form $A\vec{z} = \vec{b}$, where $\vec{z}$ consists of the position of train 2 at each time step, its velocity at each time step and its acceleration at each time step.

2.3.4 Algorithm

We ran simulations in Matlab using the built-in minimization routine `fmincon`. The `fmincon` function uses a quasi-Newton line-search method to find the minimum. Our algorithm consists of preparing inputs for `fmincon`.

Step 1. Initialization of parameters

- Time in seconds when train 2 will reach the station.
- Length of train.
- Size of each discrete time step and consequent number of grid points.

- Interference tolerance ε.
- Speed limit, maximum possible acceleration and maximum possible deceleration.
- Distance to station and train's dwell time in station.
- Initial position, velocity and acceleration of train 2.
- Position, velocity and acceleration of train 1 at each grid point.
- Initial guess vector for the minimization routine `fmincon`.
- Upper and lower bounds for each variable at each time step.

Step 2. Set up objective function and constraints

- Use the trapezoidal rule to construct a discretized cost function f.
- Construct a matrix of equality constraints (2.2)-(2.8), called A_E and corresponding vector b_E, which forces the method to follow the discretized equations of motion and the initial and terminal conditions.
- Construct a matrix of inequality constraints (2.10) and (2.11) A_I and corresponding vector b_I, which force a solution to be within the acceleration tolerance defined in constraint (2.13).
- Create an m-file for the nonlinear constraints C_I, which enforce constraint (2.12).

Step 3. Optimize the objective function

- Minimize the discretized cost function f using the function `fmincon()`.

2.4 Results

In our initial model, we changed the tolerance ε of the absolute difference in acceleration of the train at two close time steps and observed the resulting interference. We found that decreasing the tolerance resulted in a lower value for the work energy integral, and thus a lower amount of energy consumed. In Fig. 4 the graphs of acceleration versus time for smaller values of ε have shorter, smaller spikes than those with larger tolerances. By our definition, we have reduced interference and confirmed our hypothesis that minimizing energy (corresponding to lower ε values) does decrease interference. A curious feature of these graphs is the sharp drop in acceleration of train 2 occuring immediately after the simulation begins. Since there is no physical reason why the acceleration should drop to zero at first, and the size of this peak decreases with decreasing tolerance values and is unobservable in tolerances of 1 and smaller, we believe that this may be a numerical side-effect of the computation. Future investigation will hopefully provide some insight into the cause of these spikes.

We also observed the dependance of interference on other parameters, such as the length of the train, the length of the time step, and the initial starting point guess for the `fmincon` Matlab solver. We found that changing the length of the trains does not appear to decrease the overall interference of the system, although the interference pattern is much different. Changing the time step from one-half second to one second decreased interference, although this could just be due to less fine resolution. We ran the minimization program with a smaller time step of one-quarter; however, the `fmincon` function could not produce a solution after four hours, and we did not consider time steps smaller than one-half. To test the stability of our solution to the original formulation, we used the output solution for the $\varepsilon = 2.5$ as the initial guess for the problem with a larger tolerance of $\varepsilon = 20$. We assumed that with such an optimal initial guess, even though there was virtually no restriction on the rates of change of acceleration, the function output should be nearly the same. Instead, the `fmincon` function determined that there was no minimum value, contrary to our intuition. More research will be needed to understand this result.

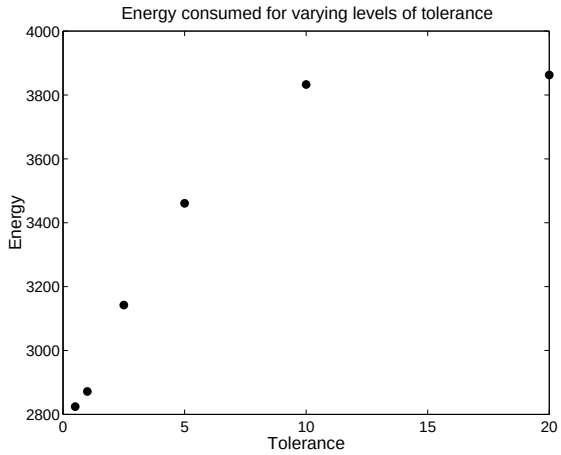


Figure 2.2: Energy consumption increases as trains are allowed to accelerate and decelerate more quickly.

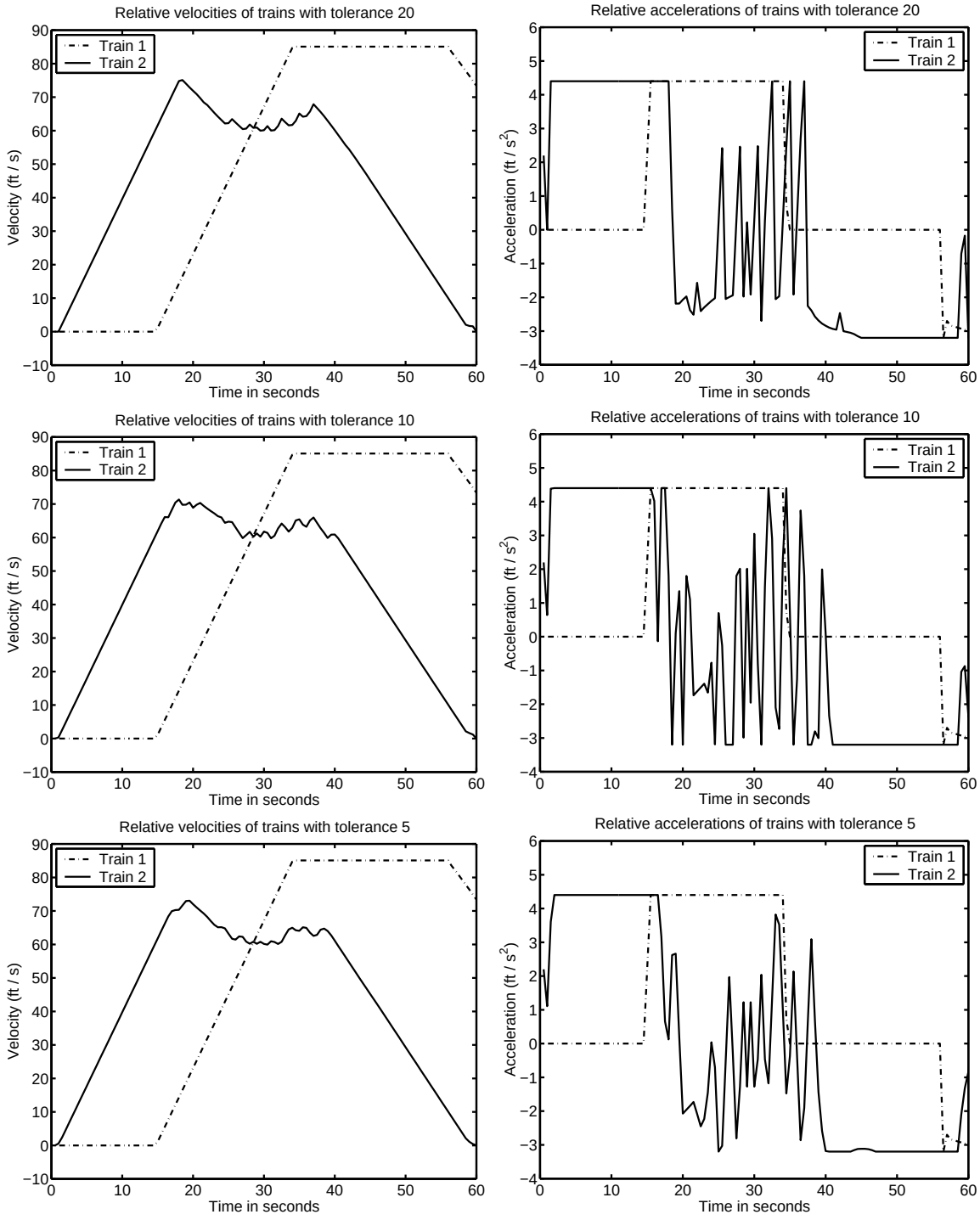


Figure 2.3: Velocity and acceleration plots for both trains with varying constraints on rate of change of acceleration: $\varepsilon = 20, \varepsilon = 10, \varepsilon = 5$.

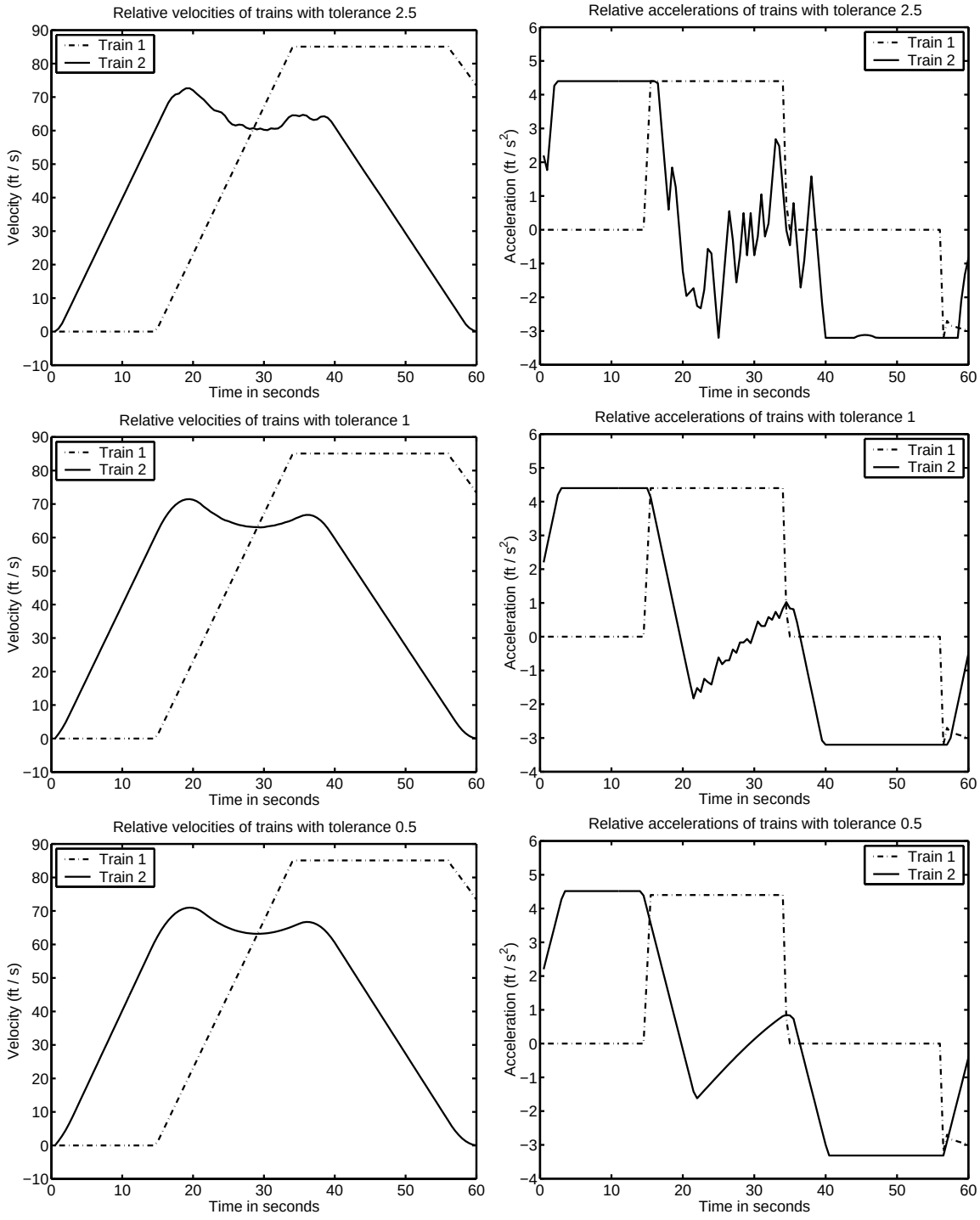


Figure 2.4: Velocity and acceleration plots for both trains with varying constraints on rate of change of acceleration: $\varepsilon = 2.5, \varepsilon = 1, \varepsilon = 0.5$.

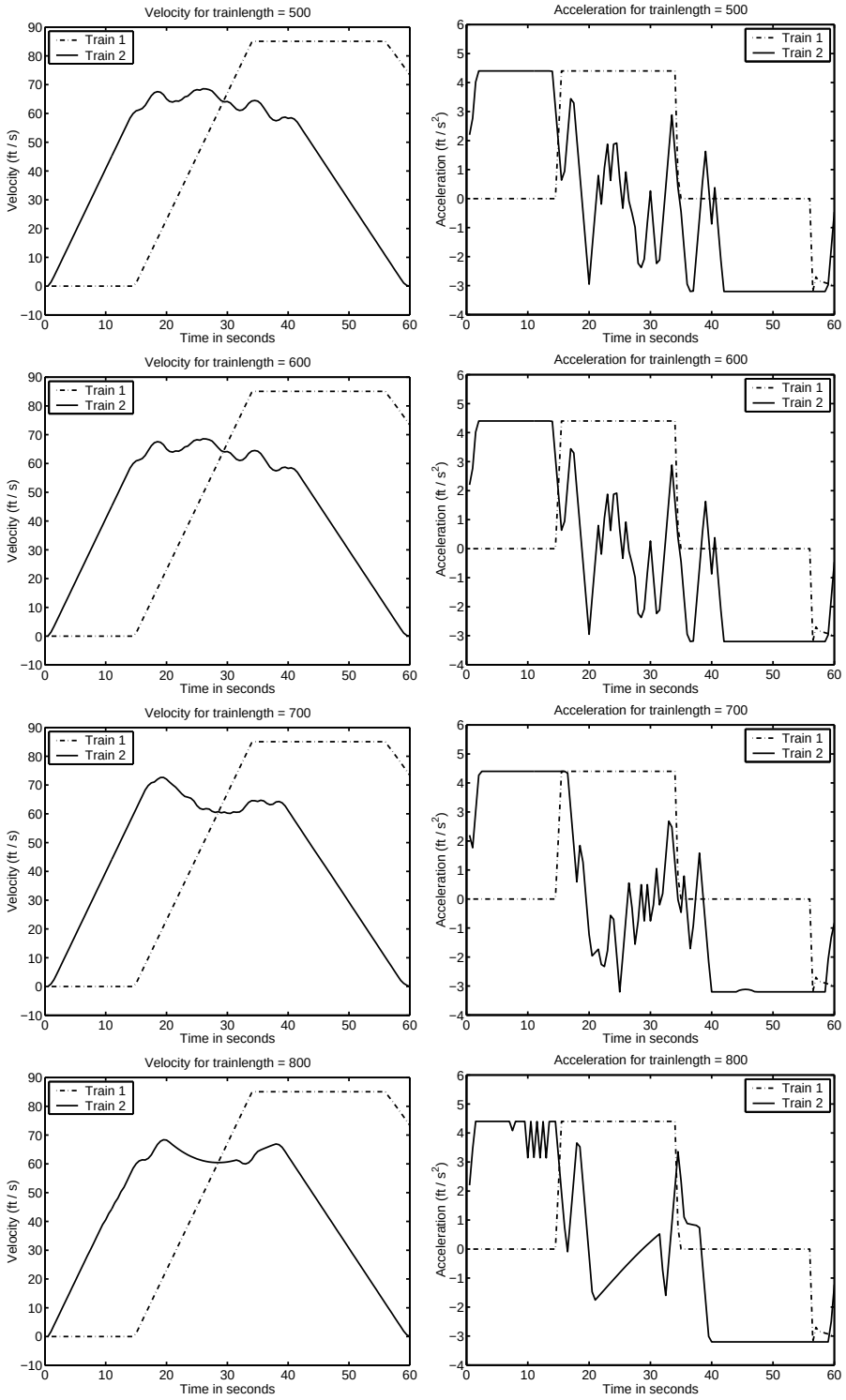


Figure 2.5: Velocity and acceleration plots for both trains with varying train lengths for tolerance $\varepsilon = 2.5$.

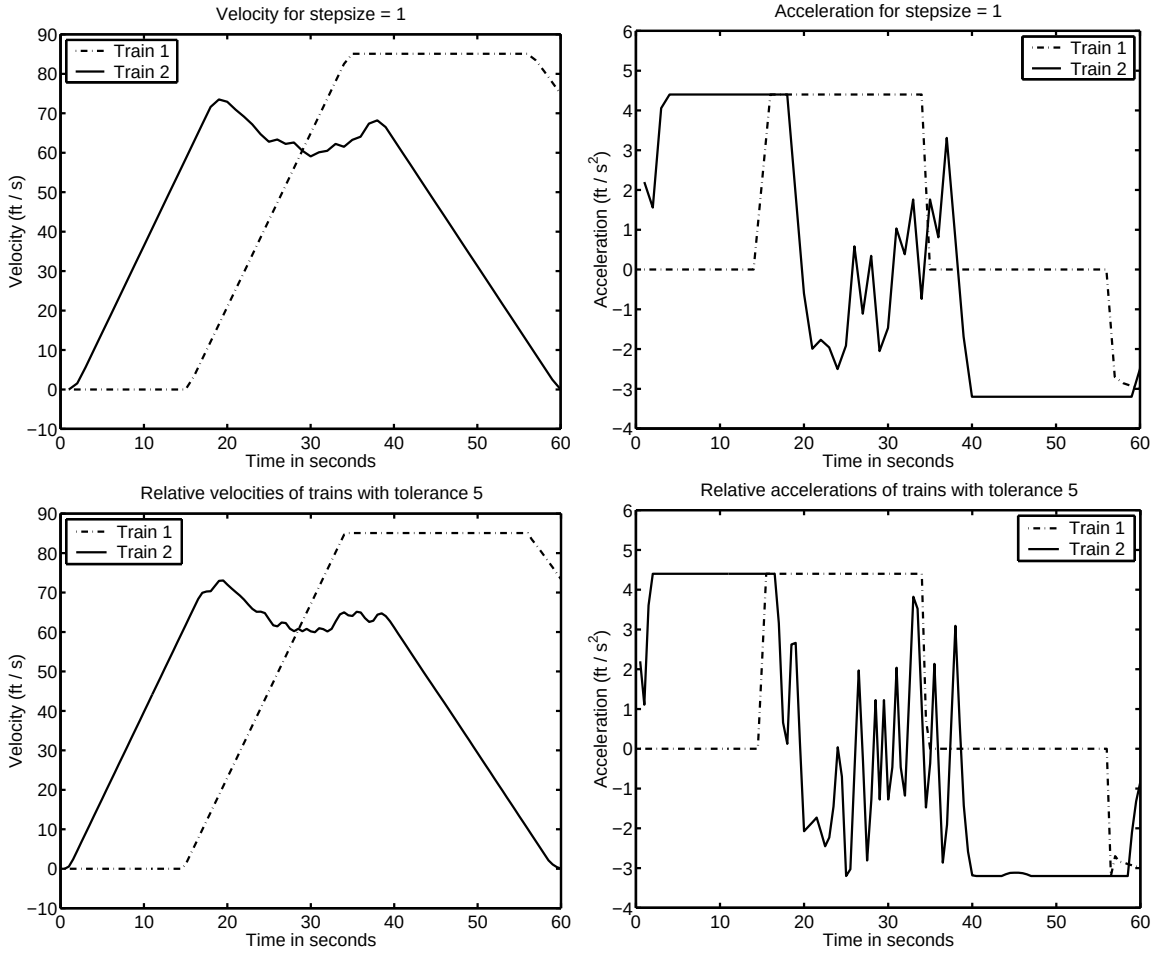


Figure 2.6: Velocity and acceleration plots step size 1 with $\varepsilon = 2.5$ and step size 1/2 with $\varepsilon = 5$.

2.5 Discussion

When we first began to address this problem, we tried to optimize a two-train, one-station system where the motion of the first train is computed during the optimization process. This approach turned out to be computationally infeasible, as it required recording the values of variables for both trains, which eventually exceeded the available memory. This problem can be avoided using a different software package with a nonlinear solver more suited to handling the sparse matrices generated, or by coding the problem in a different programming language. Making the best of our technological limitations however, we were able to utilize Matlab's nonlinear optimization solver to produce results consistent with what we had expected: train interference decreases as the rate of change in acceleration over a short time interval is constrained.

Future research will expand upon these results. The most important modification to the code we have developed would be to include more trains and stations in our model, more accurately matching the real BART train line. Adding an element of randomness in arrival time to simulate delays will be crucial if this formulation is to be applied to the real-world system. For a more accurate mathematical model, the effects of friction and air resistance must be included in our work integral cost function. In addition to the alterations we have already considered, changing the speed limit along a length of track and the scheduled stop time at stations should provide interesting variations in interference patterns.

Expanding upon our results, BART engineers will be able to further minimize train interference, drastically cutting fuel costs while increasing the level of customer satisfaction. The money saved by BART can go back into the system to make further improvements. The improved operation of the trains will encourage more people to use public transportation, increasing revenues for BART and reducing the strain on environmental resources.

Bibliography

[1] B. J. Driessen, *On-off minimum-time control with limited fuel usage: global optima via linear programming*, Albuquerque, NM 87185-0847.

[2] S. P. Gordon, *Enhanced algorithms for advanced automatic train control*, Sandia National Laboratories, DRAFT - 11/11/96.

[3] R. J. Vanderbei, *Case studies in trajectory optimization: trains, planes, and other pastimes*, Princeton University, ORFE-00-3, 2000.

Appendix

```
% FILENAME: t2g1.m
% PURPOSE: to solve the train control energy minimization problem
% Revision to train2.m
% IMMW Group 2
%
% This file assumes a given sequence of information regarding position, velocity
% and acceleration of train 1, in order to minimize energy of train 2
%
% Also needed: nonlincon2.m, cost2.m
%

clear all
timebegin=cputime;
global train1
train1=load('train1_70s.txt'); % Train 1 information loaded from file
train2=load('excel_train2.txt'); % Train 2 information loaded from file
T2=60; % T2 is the arrival time (s)for train2 at the station
T=T2; % time (s) length of simulation
small=0.0001; % Small number for numerical caluculations
imp=2.2;  % initial acceleration (ft/(s*s)) when trains start to move
global h  % time increments (one half a second)
h=1/2;

global gp % number of grid points needed for each variable
gp=T/h+1;

tol=5; % magnitude restriction on acceleration rate of change (ft/(s*s*s)),
        % related to definition of interference
z0=zeros(3*gp,1); % initial condition (in the form [x_2,v_2,a_2])
% used as the initial guess in the search algorithm fmincon
%Initial guess consists of given data points
vmax=88; % maximum speed limit (ft/s)
amin=-3.2; % minimum deceleration rate (ft/s^2)
amax=4.4; % maximum acceleration rate (ft/s^2)
L=2640;  % Distance (ft) to the last station from zero position

dwell=15; % dwell time (sec) at station

x20=0; v20=0; a20=amax/2;   % initial conditions
x2T=L; v2T=small; a2T=0;   % terminal conditions

lb=[zeros(gp,1); zeros(gp,1);amin*ones(gp,1)];      % lower bound for the vector z
ub=[L*ones(gp,1);vmax*ones(gp,1);amax*ones(gp,1)];  % upper bound vector for z

% Constraint Matrices to input into fmincon

nomat=zeros(gp-1,gp);        % nomat is block zero matrix for construction of AI,AE
AIblock=diag(ones(1,gp),0)+diag(-ones(1,gp-1),1);  AIblock=AIblock(1:gp-1,:);
AItop=[nomat,nomat,AIblock];  AI=[AItop;-AItop];
AI=sparse(AI);                      % AI is inequality constraint matrix
bI=sparse(tol*h*ones(2*(gp-1),1));       % bI is inequality vector
```

```
AEblock=-AIblock;  AEhbig=diag(-h*ones(1,gp-1),1);  AEh=AEhbig(1:gp-1,:);
AE1=[AEblock,AEh,nomat];  AE2=[nomat,AEblock,AEh];  AE=[AE1;AE2];
AE=sparse(AE);                  % AE is equality constraints
bE=sparse(zeros(2*(gp-1),1)); % bE is equality vector

clear AE1 AE2 AEhbig AEh AEblock AIblock AItop nomat

% matrix used to assign the initial and
% terminal positions (x), velocities (v), accelerations (a)
% IT is a 6 row matrix
for i=1:3
   IT(i,(i-1)*gp+1)=1;
   IT(i+3,i*gp)=1;
end

itb=[x20;v20;a20;x2T;v2T;a2T];  % initial and then terminal points

AE=[AE; IT];  bE=[bE; itb];

clear IT itb imp train2 T2 T small vmax amin amax dwell x20 v20 a20 x2T v2T a2T

[z,fval,exitflag,output,lambda,grad,hessian]=fmincon('cost2',z0,AI,bI,AE,bE,lb,ub,'nonlincon2');

timefinish=cputime-timebegin

% save tolerance2.5.mat

% Distance plots
figure
plot([1:gp]*h,train1(1:gp),'-.','LineWidth',2)
hold on
plot([1:gp]*h,z(1:gp),'LineWidth',2)
xlabel('Time in seconds','FontSize',12)
ylabel('Position along track','FontSize',12)
axis([0 60 0 6000])
title(['Relative positions of trains with tolerance ',num2str(tol)],'FontSize',12)
legend('Train 1', 'Train 2',4)
% Velocity plots
figure
plot([1:gp]*h,train1(gp+1:2*gp),'-.','LineWidth',2)
hold on
plot([1:gp]*h,z(gp+1:2*gp),'LineWidth',2)
xlabel('Time in seconds','FontSize',12)
ylabel('Velocity','FontSize',12)
axis([0 60 -10 90])
title(['Relative velocities of trains with tolerance ',num2str(tol)],'FontSize',12)
legend('Train 1', 'Train 2',4)
% Acceleration plots
figure
plot([1:gp]*h,train1(2*gp+1:3*gp),'-.','LineWidth',2)
hold on
plot([1:gp]*h,z(2*gp+1:3*gp),'LineWidth',2)
xlabel('Time in seconds','FontSize',12)
```

```
ylabel('Acceleration','FontSize',12)
axis([0 60 -4 6])
title(['Relative accelerations of trains with tolerance ',num2str(tol)],'FontSize',12)
legend('Train 1', 'Train 2',1)



function y=cost2(z)

% cost function for optimization

global h
global gp

% form a new acceleration vector
accel=[z(2*gp+1:3*gp)];

% form a new velocity vector
vel=[z(gp+1:2*gp)];

% Take the positive part of acceleration for Work
accel=.5*(accel+abs(accel));

% do Trapezoidal rule approximation of Work Integral
y=trapz(accel.*vel)*h;



function [c,ceq]=nonlincon2(z)

% function for nonlinear inequalities
% function enforces mininmum safety distance limits

global gp
global train1

for i=1:gp
c(i,1)=z(i)-train1(i)+550+(1/9)*z(gp+i)^2+8*z(gp+i);
end

ceq=0;
```

Report 3

Mathematical Modeling of Skin Paint Studies

Mathematical Modeling of Comparative Initiation/Promotion Skin Paint Studies of B6C3F$_1$ Mice and Swiss CD-1 Mice.

Nusrat Jahan[1], Ya Jin[2], Youngsuk Lee[3], Yaxi Zhao[4], and Rebekah Stephenson[5]

Industrial Problem Presenter :
Shree Whitaker, Chris Portier, Fred Parham National Institute of Environmental Health Sciences, NIH

Faculty Consultant : Hien Tran, North Carolina State University

In this study we mathematically describe a cancer initiation/continuous promotion mechanism and estimate the related biological parameters. The data on Swiss CD-1 and B6C3F$_1$ mice was collected by NTP(Design A, 1994). These mice were initiated with DMBA and then promoted with TPA on a weekly basis. By varying the dosage of DMBA and the type of mice, we analyze four different subsets of the original data. This study identifies a working model to describe the mutation of normal cells to papillomas, then the final mutation of papillomas into carcinomas for each of the subsets. Our model assumes that there are multiple stages from initiation to papilloma. For each stage of the mutation, we assume any single cell will either mutate or not. Therefore, the underlying probability distribution of the number of papillomas at the initiated stage is binomial. For similar reasons, at the final stage after promotion, the probability distribution of the number of carcinomas is also binomial. We try to ascertain a general model, which would account for the data from all four groups. Finally, we compare the cell birth rate for the papilloma model between two strains of mice for the same dosage of DMBA. We also compare the birth rates for different dosage of DMBA within each strain of mice.

[1]Mississippi State University
[2]Brown University
[3]University of Wisconsin Madison
[4]University of North Carolina Wilmington
[5]Michigan State University

3.1 Introduction

An initiation/continuous promotion study typically involves a single sub-threshold application of a carcinogen substance, followed by repeated applications of a non-carcinogen substance. This type of study is usually conducted on mice because they are a far more responsive model for skin initiation/promotion studies than other rodent species. At the same time, not all strains of mice are equally sensitive to the initiation/promotion protocol [2].

The data for this paper have been obtained from a one-year study conducted by NTP (National Toxicology Program, Study Design A in 1994). In that study, groups of 30 male and 30 female mice were administered 7,12-dimethylbenz(a) anthracene (DMBA) as an initiator treatment in the first week of the 52-week study period, followed by weekly application of 12-O-tetradecanoyl-phorbol-13-acetate(TPA) as a promoter treatment for the remaining 51 weeks. Different doses of DMBA in combination with different doses of TPA were used for three different strains of mice. For the purpose of our study, however, we use only the data on two different strains (Swiss CD-1 and $B6C3F_1$) of mice. We compare the sensitivity of Swiss CD-1 and $B6C3F_1$ mice strains in terms of the number of papillomas. We also compare different doses of DMBA (2.5 and $25.0\mu g$). Each group has the same repeated typical application of TPA (5 μg) Therefore, in our study, we have the following four groups:

Swiss CD-1	DMBA:	2.5μg	TPA:	5 μg
Swiss CD-1	DMBA:	25.0μg	TPA:	5 μg
$B6C3F_1$	DMBA:	2.5μg	TPA:	5 μg
$B6C3F_1$	DMBA:	25.0μg	TPA:	5 μg

Consistency of the data was maintained by the standard method of recording clinical observations, whereby the appearance and progression of any tumor development on the skin were recorded. When a skin tumor first appeared, it was considered a tissue mass, until it became at least 1 mm in diameter and had been present for 14 days. Then, the tissue mass was considered a papilloma. Furthermore, when a papilloma became necrotic in appearance and was attached to the underlying tissue, it was recorded as a carcinoma. In addition, microscopic evaluations were carried out to confirm the state of carcinoma.

3.2 The Model

The design of our model is two-fold, incorporating growth of papilloma and then carcinoma. First, we focus on modeling the growth of a papilloma (see Figure 3.1) and then substitute the papilloma model into the overall model of the probablility of a normal cell forming a carcinoma. In Figure 3.2, the papilloma stage is represented by "Initiated Cells." The meanings of the parameters of the model are explained below.

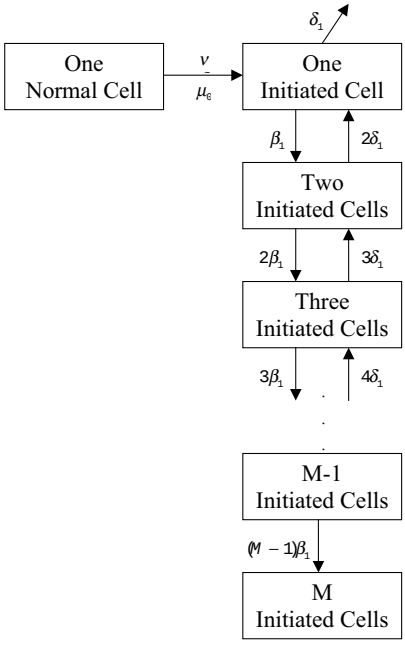


Figure 3.1: Model of a Normal Cell Generating a Papilloma.

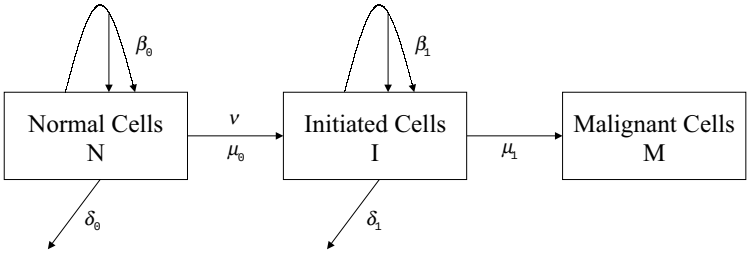


Figure 3.2: Model of a Normal Cell Generating a Carcinoma.

β_0: the birth rate of a normal cell (set equal to 0)
δ_0: the death rate of a normal cell (set equal to 0)
β_1: the birth rate of an initiated cell
δ_1: the death rate of an initiated cell
μ_0: the mutation rate of a cell from the normal to initiated state
μ_1: the mutation rate of a cell from the initiated state to malignant state
ν: the instantaneous mutation rate from the normal to initiated state
m: the number of the normal cells at the beginning for each animal
M: the minimal number of initiated cells needed to comprise a detectable papilloma

Let I_i represent the ith initiated stage; i.e., the stage in papilloma development with i initiated cells present, and let $Q_{iM}(s,t)$ represent the probability that a cell at stage I_i at time $t-s$ will not reach the stage I_M before the time t.

We now describe mathematically the model for the papilloma data (see Figure 3.1). We make the following assumptions:

Assumption 1: Initiated cells follow a linear birth-death process with constant rates.

Assumption 2: The minimal number of initiated cells, M, is large enough that we may ignore stages after I_M.

Assumption 3: One normal cell yields at most one papilloma.

Under these assumptions the papilloma stage of our two-stage model can be described by the following system of differential equations and initial conditions (see Appendix for derivations):

$$\begin{aligned}
\frac{dQ_{0M}(s,t)}{ds} &= -Q_{0M}(s,t)\mu_0 + Q_{1M}(s,t)\mu_0 && (3.1)\\
\frac{dQ_{1M}(s,t)}{ds} &= Q_{2M}(s,t)\beta_1 - Q_{1M}(s,t)(\beta_1+\delta_1) + \delta_1 && (3.2)\\
\frac{dQ_{iM}(s,t)}{ds} &= iQ_{i+1,M}(s,t)\beta_1 - iQ_{iM}(s,t)(\beta_1+\delta_1) + iQ_{i-1,M}(s,t)\delta_1 && (3.3)\\
Q_{MM}(s,t) &= 0, \qquad \forall s,t && (3.4)\\
Q_{iM}(0,t) &= 1, \qquad i=0,1,\ldots,M-1. && (3.5)
\end{aligned}$$

The system (3.2) – (3.5) can be solved analytically [8], and the solution can be written as

$$Q_{1M}(s,t) = \begin{cases} 1 - \dfrac{(\beta_1-\delta_1)\beta_1^{M-1}\left(1-e^{-(\beta_1-\delta_1)s}\right)^{M-1}}{\left[\beta_1-\delta_1 e^{-(\beta_1-\delta_1)s}\right]^M}, & \delta_1 \neq \beta_1 \\ 1 - \dfrac{(\beta_1 s)^{M-1}}{(1+\beta_1 s)^M}, & \delta_1 = \beta_1 \end{cases}$$

Now we use equation (3.1) to solve for $Q_{0M}(s,t)$.

Figure 3.1 illustrates that, when starting from m normal cells, the number of papillomas has a binomial distribution since each normal cell can either evolve into a papilloma or not. Now consider the incidence of papilloma over time.

$$\begin{aligned}
P_{NP}(t) &\doteq P(\text{1 normal cell reaching stage } I_M \text{ (papilloma) before } t, \text{ starting at time 0})\\
&= P(\text{1 normal cell} \to \text{papilloma} \,|\, \text{no mutation at } t=0)\\
&\qquad \times P(\text{no mutation at time 0})\\
&+ P(\text{1 normal cell} \to \text{papilloma} \,|\, \text{mutation at } t=0) \times P(\text{mutation at time 0})\\
&= [1-Q_{0M}(t,t)](1-\nu) + [1-Q_{1M}(t,t)]\nu
\end{aligned}$$

For a particular group and a particular mouse, let $X(t)$ be the number of papilloma before t, starting with m normal cells at time 0. Then

$$P[X(t)=x] = \binom{m}{x}(P_{NP}(t))^x(1-P_{NP}(t))^{m-x}, \qquad x=0,1,2,...,m,$$

and

$$\mathrm{E}[X(t)] = mP_{NP}(t) = m([1-Q_{0M}(t,t)](1-\nu) + [1-Q_{1M}(t,t)]\nu).$$

For the carcinoma incidence analysis, we have to treat the two-stage model as one system (see Figure 3.2). To describe the system, two ordinary differential equations are needed. Let $P_{02}(s,t)$ denote the probability of one normal cell not reaching carcinoma before time t, starting at time $t-s$ and $P_{12}(s,t)$ the probability of one initiated cell not reaching carcinoma before time t, starting at time $t-s$. Then

$$\begin{aligned}
\frac{dP_{02}(s,t)}{ds} &= \beta_0 P_{02}(s,t)^2 + \delta_0 + \mu_0 P_{12}(s,t) - (\beta_0+\delta_0+\mu_0)P_{02}(s,t)\\
\frac{dP_{12}(s,t)}{ds} &= \beta_1 P_{12}(s,t)^2 + \delta_1 + \mu_1 P_{12}(s,t)P_{22}(s,t) - (\beta_1+\delta_1+\mu_1)P_{12}(s,t)
\end{aligned}$$

Similarly to the papilloma stage, we have several conditions:

$$\beta_0 = \delta_0 = 0, \quad P_{02}(0,t) = P_{12}(0,t) = 1, \quad P_{22}(s,t) = 0.$$

After simplification, we obtain

$$\begin{aligned}\frac{dP_{02}(s,t)}{ds} &= \mu_0 P_{12}(s,t) - \mu_0 P_{02}(s,t)\\ \frac{dP_{12}(s,t)}{ds} &= \beta_1 P_{12}(s,t)^2 + \delta_1 - (\beta_1 + \delta_1 + \mu_1) P_{12}(s,t).\end{aligned}$$

For a certain group, a certain mouse, consider the random variable $Y(t)$ define by $Y(t) = 1$, if carcinoma is detected before time t starting from m normal cells at time 0; $Y(t) = 0$, otherwise. By the definition of $P_{12}(t,t), P_{02}(t,t)$ and ν, we know that the probability of one normal cell not reaching carcinoma before time t, starting from time 0 is $\nu P_{12}(t,t) + (1-\nu)P_{02}(t,t)$. So $P[Y(t) = 1] = 1 - [\nu P_{12}(t,t) + (1-\nu)P_{02}(t,t)]^m$, since we assume m normal cells act independently.

We now use the method of maximum likelihood to derive estimators for the parameters of our model.

1. The Papilloma Stage

Let x_{ijk} represent the number of papillomas for the $i-$th animal in the $j-$th experimental group at the time k. The likelihood function for the number of papillomas can then be expressed as

$$L_1 = \prod_i \prod_j \prod_k P\left[X(t) = x_{ijk}\right]. \tag{3.6}$$

Taking the natural logarithm of (3.6) yields

$$\ln L_1 = \sum_i \sum_j \sum_k \ln \left[\binom{m}{x_{ijk}} P_{NP}^{x_{ijk}} (1 - P_{NP})^{(m - x_{ijk})} \right]$$

Since we are maximizing L_1, or equivalently $\ln(L_1)$, with respect to P_{NP}, the constant term can be ignored, leaving

$$\sum_i \sum_j \sum_k [x_{ijk} \ln P_{NP} + (m - x_{ijk}) \log(1 - P_{NP})].$$

2. The Carcinoma Stage

Let y_{jk} represent the number of malignant tumors in the $j-$th experimental group at the time k. The likelihood function is then defined as

$$L_2 = \prod_j \prod_k P\left[Y(t) = y_{jk}\right]$$

having corresponding natural logarithm

$$\ln L_2 = \sum_j \sum_k \ln P\left[Y(t) = y_{jk}\right]$$

To attain estimators $\{\hat{\beta}_1, \hat{\delta}_1, \hat{\mu}_0, \hat{\mu}_1, \hat{\nu}\}$, the function $\ln L_1 + \ln L_2$ is maximized over all possible values of $\{\beta_1, \delta_1, \mu_0, \mu_1, \nu\}$.

After achieving the optimally estimated parameters, we can calculate the incidence of papilloma $P_{NP}(t)$, expected number of papilloma $E[X(t)]$, the incidence of carcinoma $1 - P_{02}(t,t)^m$, etc. Then we can compare the difference of all these values among different groups.(i.e. different initiators, promotors, doses, strains, etc.)

3. Likelihood Ratio Test

The likelihood ratio test statistic is used for testing the null-hypothesis $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. The corresponding statistic is

$$\lambda(x) = \frac{L(\hat{\theta_0} \,|\, x)}{L(\hat{\theta} \,|\, x)}$$

where x is the data, $\hat{\theta}_0 = \hat{\theta}_0(x)$ is obtained by maximizing $L(\theta|x)$ over the parameter subspace Θ_0 and $\hat{\theta} = \hat{\theta}(x)$ is obtained by maximizing $L(\theta|x)$ over the whole parameter space Θ.

The asymtotic distribution of the statistic $-2\log\lambda(x)$ is a χ^2 distribution with degrees of freedom being the difference in number of parameters of the two hypotheses.[1].

We apply this theory to test the hypothesis that the cell birth rates between the two strains of mice in our study are equal. Thus for our problem

$$H_0 : \beta_1^s = \beta_1^b, \quad H_1 : H_0 \text{ is not true,}$$

where the superscript s stands for Swiss CD-1 and b for B6C3F$_1$. These hypotheses yield $\Theta_0 = \{\beta_1, \delta_1^s, \mu_0^s, \nu^s, \delta_1^b, \mu_0^b, \nu^b\}$, and $\Theta = \{\beta_1^s, \delta_1^s, \mu_0^s, \nu^s, \beta_1^b, \delta_1^b, \mu_0^b, \nu^b\}$, and therefore

$$\log\lambda(x) = \left[\log L(\hat{\theta_0}|x) - \log L(\hat{\theta}|x)\right].$$

The value of $-2\log\lambda(x)$ is compared with $\chi^2(1)$. At 0.05 level of significance, the null hypothesis will be rejected if $-2\log\lambda(x) > \chi^2(1) = 3.84$. In this case, the conclusion will be that the birth rates of initiated cells are not same for the two different strains of mice. The results are summarized in Section 3.3.

3.3 Results

The method of maximum likelihood method was used to obtain expressions for estimators of the biological parameters of the initiation/promotion model of skin cancer. The initial values for the parameters were obtained from the work of Kopp-Schneider and C.J. Portier [4]. In their work, they found that the cell-cycle time of an initiated cell with promotion is 20 hours. Since our data are the numbers of papillomas per week for each mouse, the initial values for the biological parameters of an initiated cell with promotion translate into In our model we considered the number of normal cells to be $m = 12 * 10^6$ and number of initiated

$\beta_0 = 10$ births/week
$\delta_0 = 10$ deaths/week
$\mu_0 = 1$ mutation/week
$\nu_1 =$ probability of instantaneous mutation
$\beta_1 = 10$ births/week
$\delta_1 = 10$ deaths/week
$\mu_1 = 1$ mutation/week

Table 3.1: B6C3F$_1$ Mice.

	2.5 μg DMBA	25.0 μg DMBA
$\hat{\mu}_1^b$	$1.22 * 10^{-6}$	$1.9620 * 10^{-3}$
$\hat{\beta}_1^b$	3.933	3.9075
$\hat{\delta}_1^b$	3.9463	3.9830
$\hat{\nu}_1^b$	$3.5584 * 10^{-4}$	$1.2185 * 10^{-2}$

Table 3.2: Swiss CD-1 Mice.

	2.5 μg DMBA	25.0 μg DMBA
$\hat{\mu}_1^s$	0.0713	0.0275
$\hat{\beta}_1^s$	5.922	4.1881
$\hat{\delta}_1^s$	5.2961	4.2696
$\hat{\nu}_1^s$	0.1015	0.8059

cells needed to form a visible papilloma to be $M = 387$ [3]. Maximum likelihood estimates for the parameters related to B6C3F$_1$ and Swiss CD-1 mice are presented in Tables 3.1 and 3.2 respectively.

Analysis of the Results

1. Figures 3.3 and 3.4 depict the graphs of the empirical average (observed) and expected numbers of papillomas (under the model) for B6C3F$_1$ mice with DMBA dosage of 2.5 and 25.0 μg respectively. For DMBA dosage of 2.5 μg, the fit appears to be reasonably good after the 27-th week (Fig. 3.3). The poor fit in the early stages of the study may be due to the assumption of zero birth and death rate of normal cells in our model. The fit appears to be quite good for DMBA dosage of 25.0 μg (Fig. 3.4).

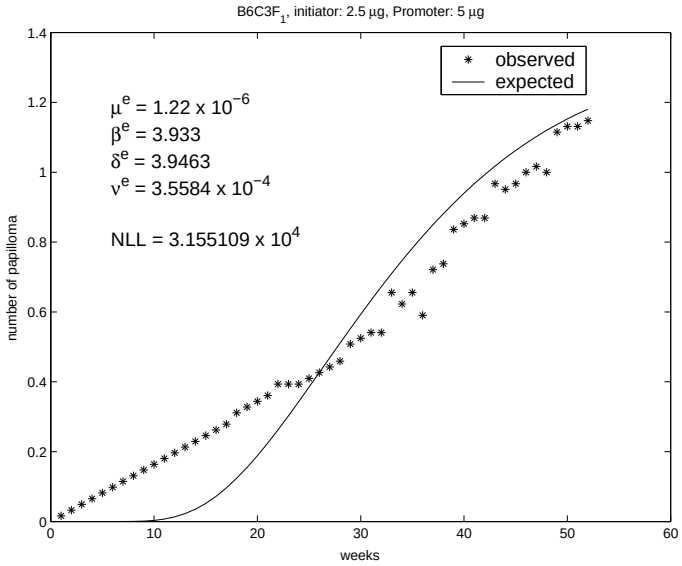


Figure 3.3: Model fit to the papilloma count for B6C3F$_1$ mice initiated with 2.5 μg DMBA.

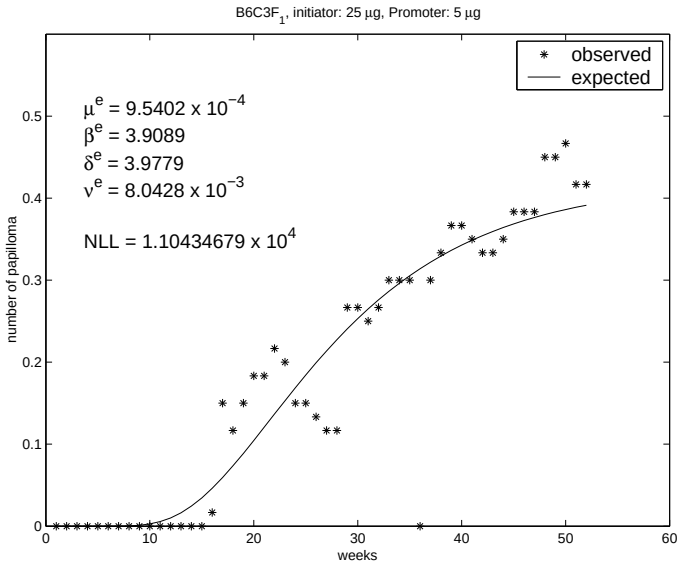


Figure 3.4: Model fit to the papilloma count for B6C3F$_1$ mice initiated with 25.0 μg DMBA.

2. Figures 3.5 and 3.6 depict the graphs of the empirical average and expected numbers of papillomas (under our model) for Swiss CD-1 mice with DMBA dosage of 2.5 and 25.0 μg respectively. The fit appears to be quite good for both DMBA dosages of 2.5 μg (Fig. 3.5) and 25.0 μg (Fig. 3.6).

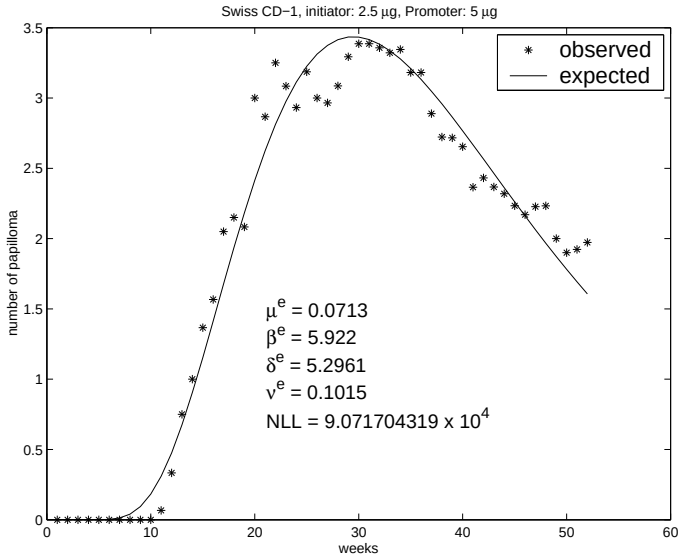


Figure 3.5: Model fit to the papilloma count for Swiss CD-1 mice initiated with 2.5 μg DMBA.

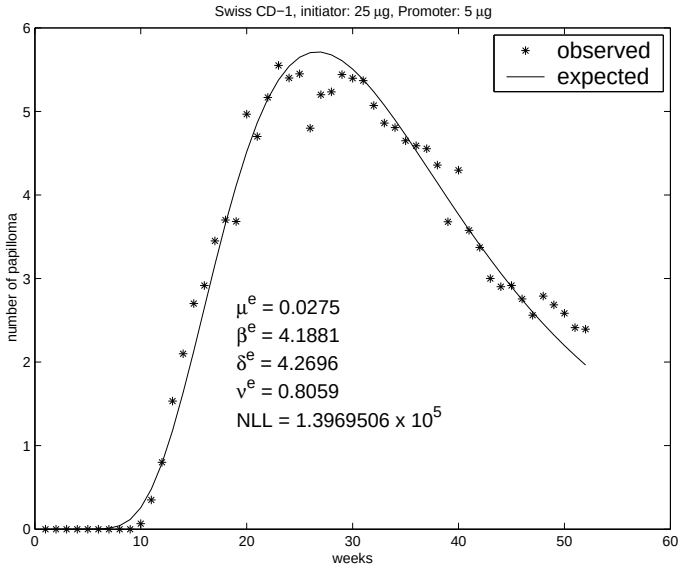


Figure 3.6: Model fit to the papilloma count for Swiss CD-1 mice initiated with 25.0 μg DMBA.

3. In order to find out whether the two different strains of mice are significantly different with respect to their birth rates, we use the likelihood ratio test statistic [1] to test the following hypotheses:

H_0: Swiss CD-1 and B6C3F$_1$ mice have equal birth rates.
H_1: Swiss CD-1 and B6C3F$_1$ mice have unequal birth rates.

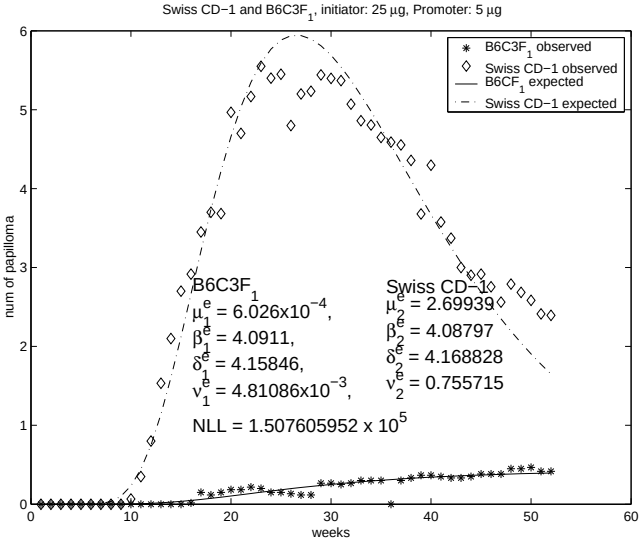


Figure 3.7: Model fit to the papilloma count for Swiss CD-1 and B6C3F$_1$ mice assuming different birth rates (both groups were initiated by 25.0 μg DMBA).

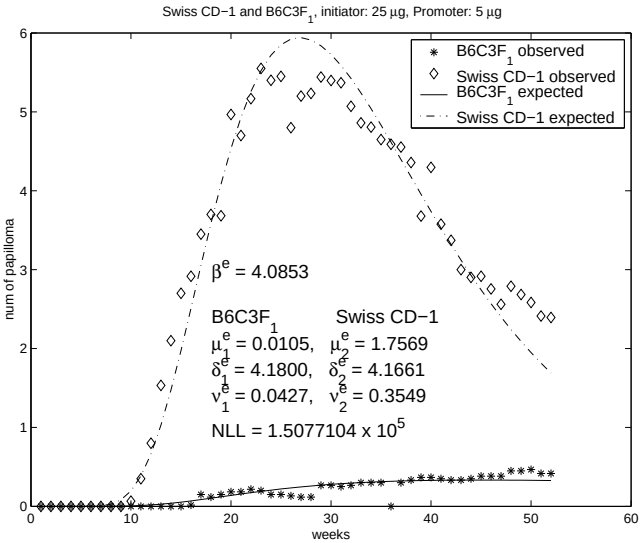


Figure 3.8: Model fit to the papilloma count for Swiss CD-1 and B6C3F$_1$ mice assuming equal birth rates (both groups were initiated by 25.0 μg DMBA).

From Figures 3.7 and 3.8, the difference in birth rates is apparent. Moreover the difference in likelihoods is 10.448 ($1.5077104 \times 10^5 - 1.507605952 \times 10^5$), is significant having a p-value of 0.0012. Therefore, the null hypothesis is rejected. We can conclude that the Swiss CD-1 and B6C3F$_1$ mice have unequal birth rates.

4. To find out whether different levels of initiation dosage have any effect on the same strain of mice, we test the following hypotheses

   H_0: Equal birth rates for DMBA dosages of 2.5 and 25.0 μg in Swiss CD-1 mice
   H_1: Unequal birth rates for DMBA dosages of 2.5 and 25.0 μg in Swiss CD-1 mice

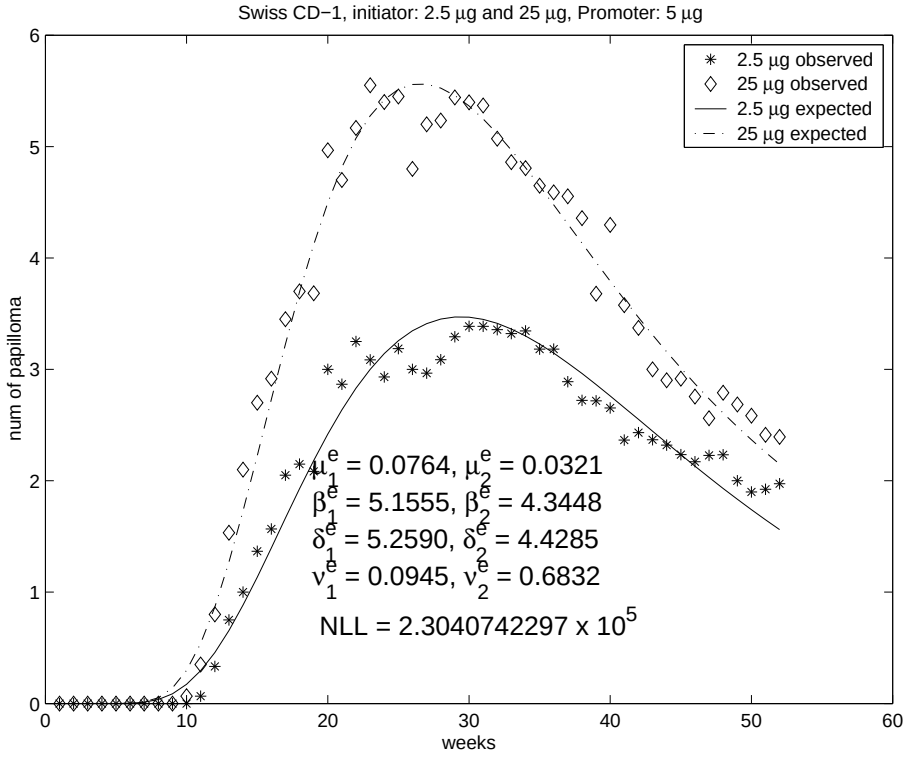


Figure 3.9: Model fit to the papilloma count for Swiss CD-1 mice initiated by 2.5 and 25 μg DMBA (assuming different birth rates for different dosages).

From Figures 3.9 and 3.10 (next page), it is apparent that the birth rates are different. The observed difference in likelihoods under two hypotheses is 55.5981 ($2.30463021 \times 10^5 - 2.3040742297 \times 10^5$), which is significant having a p-value that is less than 0.0001. Therefore, we reject the H_0 and conclude that different initiator dosages produce different birth rates for the Swiss CD-1 mice.

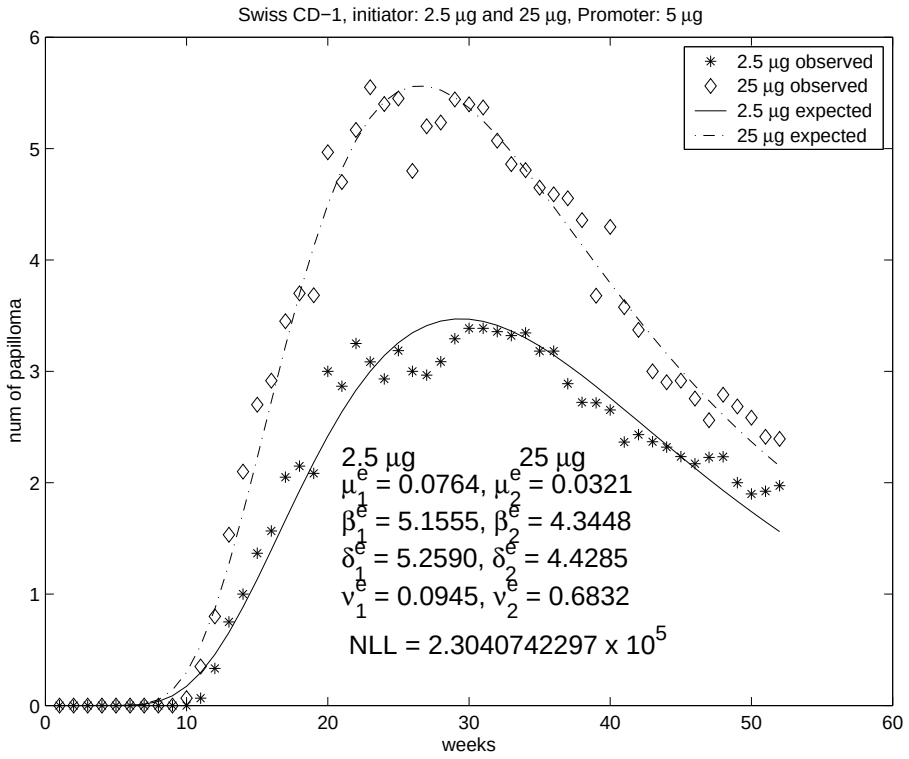


Figure 3.10: Model fit to the papilloma count for Swiss CD-1 mice initiated by 2.5 and 25 μg DMBA (assuming equal birth rates for different dosages).

3.4 Discussion

In this study we identify a general model which describes the mutation of normal cells to papilloma. We studied two different strains of mice under two separate initiator dosages. We present 4 different working models for these four cases. The underlying model is the same for all four cases based on the binomial distribution likelihood function. It should be noted that in all of these four cases the promotor and its dosage is the same (TPA 5 μg). The estimated values for the parameters tested significantly different for all four cases. Specifically we have tested for the equality of birth rate between two different strains of mice, both strains were initiated with 25 μg of DMBA. It is observed that Swiss CD-1 and B6C3F$_1$ mice have different birth rates even when the initiator dosage is the same. We also try to determine whether initiator dosage affects birth rates. In particular for Swiss CD-1 mice, we performed the test for the equality of birth rates under two different initiator dosages. Birth rates were found to be unequal for Swiss CD-1 mice for 2.5 and 25 μg DMBA.

The carcinoma data available to us were in the form of a set of summary statistics, giving insufficient information about carcinoma to conduct a meaningful maximum likelihood analysis. Therefore our likelihood function is only based on papilloma data. Once we get relevant carcinoma data, the likelihood function can be directly applied to the data, since we have derived the incidence of the carcinoma in our model. One advantage of our model is that we assume the numbers of papilloma and carcinoma have binomial distributions instead of Poisson distributions. This allows us to describe the process more accurately. Based on our model, it will be very easy to expand to a multi-stage model using a multinomial distribution. This approach should fit the data better, since it takes into account the different stages a normal cell goes through to reach the stage of papilloma. From the biological viewpoint, since there always exists some uncertainty, one may try to describe the process with a system of stochastic differential equations or with a continuous time Markov chain.

Overall, it appears that our model works quite well in the present setting and can be applied to a more general situation. However, it may be of interest to add more parameters in the model in order to account for

inherent biological complexities, such as cellular interactions, regression of papillomas etc.

Bibliography

[1] Casella, G. and Berger, R.L., Statistical Inference, Duxbury, 346-381, 1990.

[2] Comparative Initiation/Promotion Skin Paint Studies of B6C3F$_1$ Mice, Swiss (CD-1) Mice, and SENCAR Mice. Technical Report Series, No. 441, NIH Publication No. 96-3357, February 1996.

[3] Kopp-Schneider, A. Birth-death processes with piecewise constant rates. *Statist. Prob. Lett.*, **13**, 121-127, 1992.

[4] Kopp-Schneider, A. and Portier, C.J. Birth and death/differentiation rates of papillomas in mouse skin. *Carcinogenesis*, **13**, 973-978, 1992.

[5] Kopp-Schneider, A. and Portier, C.J. Carcinoma formation in mouse skin painting studies is a process suggesting greater than two-stages. *Carcinogenesis*, **16**, 53-59, 1995.

[6] S.H. Moolgavkar, D. Krewski, M.J. Goddard, A. Dewanji Two stage model for carcinogenesis: number and size distributions of premalignant clones in longitudinal studies. *Mathematical Biosciences*, **155**, 1-12, 1999.

[7] C.J. Portier, A. Kopp-Schneider, C.D. Sherman Multistage, stochastic models of the cancer process: a general theory for calculating tumor incidence. *Stochastic Environmental Research and Risk Assessment*, **14**, 173-179, 2000.

[8] Smith, Marjo V. and Portier, C.J. Incorporating observability thresholds of tumors into the two-stage carcinogenesis model. *Mathematical Biosciences*, **163**, 75-89, 2000.

Appendix

The system of ordinary differential equations derived in this section represents the probability of a normal cell forming a papilloma. Correspondingly, this derivation does not include the malignant stage. The derivation of the ordinary differential equations for the probabilities of two-stage mutation of normal cells into malignant cells has been done previously by Marjo V. Smith and Christopher J. Portier [8]. We apply this tecnique to the probability of a normal cell forming a papilloma. The difference is that we consider forming a papilloma to be a multistage process (as described in Section 3.2) and thus the result of our derivation is a system of differential equations rather than a single equation. As done by Smith and Portier, we will assume that cells act independently and set the normal-cell birth rate (β_0) and death rate (δ_0) to zero (because the size of the initial sample remains constant). In the notation introduced in Section 3.2, the following events may happen over a time interval $[t - s - \Delta s, t - s]$:

1. A normal cell mutates with probability $\Delta s\, \mu_0$.
2. A normal cell does not change with probability $1 - \Delta s(\beta_0 + \delta_0 + \mu_0)$.
3. An initiated cell replicates with probability $\Delta s \beta_1$.
4. An initiated cell dies with probability $\Delta s \delta_1$.
5. An initiated cell does not change with probability $1 - \Delta s(\beta_1 + \delta_1)$.

We first derive the equations for two special cases, $Q_{0M}(s,t)$ and $Q_{1M}(s,t)$, followed by the general case, $Q_{iM}(s,t), i = 2, 3, \ldots, M-1$.

3.4.1 The Equation for $Q_{0M}(s,t)$

Following [8], there are only four events that may happen to a single normal cell over the interval $[t-s-\Delta s, t-s]$:

- nothing may happen, so there is still one normal cell at the time $t-s$;
- the normal cell may replicate, so that there are two normal cells at time $t-s$;
- the normal cell may die, so the probability of no papilloma is 1;
- the normal cell may mutate, so the stage I_1 is achieved.

Thus we have:

$$\begin{aligned}
Q_{0M}(s+\Delta s,t) &= P(\text{no papilloma is visible at } t \,|\, \text{one normal cell } I_0 \text{ at } t-s-\Delta s) \\
&= P(\text{no papilloma is visible at } t \,|\, \text{one normal cell } I_0 \text{ at } t-s) \\
&\quad \times P(\text{1 normal cell } I_0 \text{ at } t-s \,|\, \text{one normal cell } I_0 \text{ at } t-s-\Delta s) \\
&\quad + P(\text{no papilloma is visible at } t \,|\, \text{two normal cells } I_0 \text{ at } t-s) \\
&\quad \times P(\text{two normal cells } I_0 \text{ at } t-s \,|\, \text{one normal cell } I_0 \text{ at } t-s-\Delta s) \\
&\quad + P(\text{no papilloma is visible at } t \,|\, \text{no normal cells } I_0 \text{ at } t-s) \\
&\quad \times P(\text{no normal cells } I_0 \text{ at } t-s \,|\, \text{one normal cell } I_0 \text{ at } t-s-\Delta s) \\
&\quad + P(\text{no papilloma is visible at } t \,|\, \text{no normal cells } I_1 \text{ at } t-s) \\
&\quad \times P(\text{no normal cells } I_1 \text{ at } t-s \,|\, \text{one normal cell } I_0 \text{ at } t-s-\Delta s) \\
&= Q_{0M}(s,t)[1-\Delta s(\beta_0+\delta_0+\mu_0)] \\
&\quad + (Q_{0M}(s,t))^2 \Delta s \beta_0 + \Delta\delta_0 \cdot 1 + Q_{1M}(s,t)\Delta s\, \mu_0
\end{aligned}$$

Subtracting $Q_{0M}(s,t)$ from both sides, dividing by Δs, and taking the limit as $\Delta s \to 0$, we obtain

$$\frac{dQ_{0M}(s,t)}{ds} = (Q_{0M}(s,t))^2\beta_0 - Q_{0M}(s,t)(\beta_0+\delta_0+\mu_0) + \delta_0 + Q_{1M}(s,t)\mu_0. \tag{3.7}$$

Since $\beta_0 = \delta_0 = 0$, equation 3.7 becomes

$$\frac{dQ_{0M}(s,t)}{ds} = -Q_{0M}(s,t)\mu_0 + Q_{1M}(s,t)\mu_0. \tag{3.8}$$

3.4.2 The equations for $Q_{1M}(s,t)$

In this case, there are only three events that may happen to a single initiated cell at the stage I_1 over the interval $[t-s-\Delta s, t-s]$:

- nothing may happen, so there is still one initiated cell at the time $t-s$;
- the initiated cell may replicate, so that there are two initiated cells at time $t-s$;
- the initiated cell may die, so the probability of no papilloma is 1.

Thus, we have:

$$\begin{aligned} Q_{1M}(s+\Delta s,t) &= P(\text{no papilloma is visible at } t \,|\, I_1 \text{ at } t-s-\Delta s) \\ &= P(\text{no papilloma is visible at } t \,|\, I_1 \text{ at } t-s) \\ &\quad \times P(I_1 \text{ at } t-s \,|\, I_1 \text{ at } t-s-\Delta s) \\ &\quad + P(\text{no papilloma is visible at } t \,|\, I_2 \text{ at } t-s) \\ &\quad \times P(I_2 \text{ at } t-s \,|\, I_1 \text{ at } t-s-\Delta s) \\ &\quad + P(\text{no papilloma is visible at } t \,|\, \text{no cells } I_0 \text{ at } t-s) \\ &\quad \times P(\text{no cells } I_0 \text{ at } t-s \,|\, I_1 \text{ at } t-s-\Delta s) \\ &= Q_{1M}(s,t)(1-\Delta s(\beta_1+\delta_1)) + Q_{2M}(s,t)\Delta s\beta_1 + \Delta s\delta_1 \cdot 1. \end{aligned}$$

Subtracting $Q_{1M}(s,t)$ from both sides, dividing by Δs, and taking the limit as $\Delta s \to 0$, we obtain:

$$\frac{dQ_{1M}(s,t)}{ds} = Q_{2M}(s,t)\beta_1 - Q_{1M}(s,t)(\beta_1+\delta_1) + \delta_1. \tag{3.9}$$

3.4.3 The Equations for $Q_{iM}(s,t),\ i = 2,3,...,M-1$

In this case, there are only three events that may happen to the initiated cells at the stage I_i over the interval $[t-s-\Delta s, t-s]$:

- nothing may happen, so there are still i initiated cells at the time $t-s$;
- any one of the i initiated cells may replicate, so that there are $i+1$ initiated cells at the time $t-s$ and the stage I_{i+1} is achieved;
- any one of the i initiated cell may die, so that there are $i-1$ initiated cells at the time $t-s$ and the process returns to the stage I_{i-1}.

Thus, we have:

$$\begin{aligned} Q_{iM}(s+\Delta s,t) &= Q_{iM}(s,t)(1-i\,\Delta s(\beta_1+\delta_1)) \\ &\quad + iQ_{i+1,M}(s,t)\Delta s\beta_1 + iQ_{i-1,M}(s,t)\Delta s\,\delta_1. \end{aligned}$$

Subtracting $Q_{iM}(s,t)$ from both sides, dividing by Δs, and taking the limit as $\Delta s \to 0$, we obtain:

$$\frac{dQ_{iM}(s,t)}{ds} = iQ_{i+1,M}(s,t)\beta_1 - iQ_{iM}(s,t)(\beta_1+\delta_1) + iQ_{i-1,M}(s,t)\delta_1. \tag{3.10}$$

Finally, to complete the system (3.8)-(3.10), we need initial conditions for $Q_{iM}(s,t)$. Since we consider the papilloma stage irreversible, we have

$$Q_{MM}(s,t) = 0, \quad \text{for all } s,t,$$

and, by definition,

$$Q_{iM}(0,t) = 1, \quad i = 0,1,\ldots,M-1.$$

Report 4

Mathematical Models for Articular Cartilage

Mathematical Models for Articular Cartilage: Molecular Diffusion in Photobleaching Experiments and Signal Transmission in a Chondron

Stacy Beun[1], Praveen Chaturvedi[2], Bastian Gebauer[3], Joseph Latulippe[4], Ming Zhao[5],

Problem Presenter:
Farshid Guilak
Orthopaedic Research Laboratories
Dept. of Surgery, Duke University Medical Center

Faculty Consultant:
Mansoor Haider

Abstract

We consider two mathematical models that are relevant to current experimental studies of articular cartilage that aim to understand: (i) the effect of matrix anisotropy on molecular diffusion, and (ii) the effect of pericellular matrix permeability on mechanotransduction in the cartilage cell-matrix unit, called a chondron. For the first model, we develop diffusion models for photobleaching experiments with a circular region of bleaching. Our simulations indicate that a continuous bleaching model predicts sharp fluorescence images that increase the ability to detect matrix anisotropy via image analysis. For the second model, we analyze the effect of permeability of the pericellular matrix on signal transmission in a chondron using a finite difference model in a parametric analysis. Our simulations indicate that 80-100*% signal transmission is achieved when the PCM is much stiffer than the cell and the PCM permeability is two orders of magnitude larger than that of the cell.*

4.1 Introduction

Articular cartilage is a hydrated biological soft tissue that lines the surfaces of diarthroidal joints such as the knee, shoulder and hip. The primary function of cartilage is to distribute stresses in load-bearing and to provide a low-friction surface for joint motion. While cartilage can perform these functions over a lifetime, the

[1]North Carolina State University
[2]Southern Methodist University
[3]University of Mainz, Germany
[4]Montana State University
[5]State University of New York at Stoney Brook

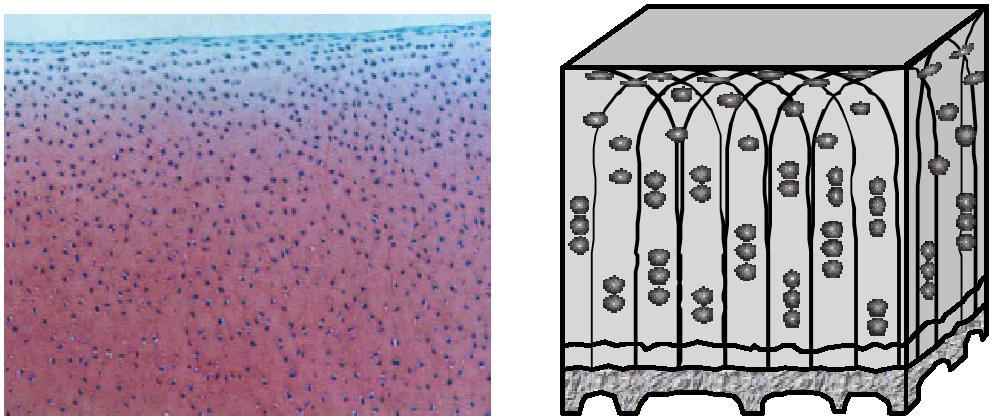


Figure 4.1: (a) A layer of articular cartilage showing chondrocytes embedded throughout the extracellular matrix. (b) A schematic of the variation of extracellular matrix anisotropy with depth in a cartilage layer

degeneration of cartilage, called osteoarthritis (OA), is a widespread condition that progresses with age. The structure of cartilage arises from an extracellular matrix (ECM) of cross-linked collagen fibers and entrapped proteoglycan macromolecules. In the surface zone of cartilage, the ECM fibers are oriented parallel to the layer surface while in the deep zone they align parallel to the subchondral bone. In the mid-zone, the fibers are isotropic (Fig. 4.1b).

Embedded in the ECM are specialized cells, called chondrocytes, whose metabolic response dictates maintenance and turnover of the ECM (Fig. 4.1a). Cartilage is avascular and aneural. Consequently, chondrocyte metabolism depends not only on inherent genetic and biochemical factors but also on mechanical and physico-chemical factors in the local cell environment. These factors include stress and ionic charge density of the ECM, pressurization of the interstitial fluid, and molecular diffusion of nutrients through the ECM to the cells. An important component of the local cell environment is the pericellular matrix (PCM), which encapsulates the chondrocyte and is believed to regulate transmission of mechanical signals to the cell. In contrast to the ECM, which is dominated by type-II collagen, the predominant collagen type in the PCM is type-VI. Together, the cell and PCM are termed a chondron. As a joint undergoes loading, mechanical signals are transmitted via the ECM to each chondron and, via the PCM, to each cell which, in response, can alter its metabolic activity. It is believed that the functional role of the PCM is to protect the cell from excessive load while, simultaneously, facilitating the transmission of mechanical signals from the ECM to the cell. A key physiological question is what components of the local mechanical and physico-chemical environment the cell uses to detect changes, and hence alter its metabolic activity.

One of the goals of the Orthopaedic Research Lab is to study the causes of osteoarthritis (OA) and the factors that influence the degenerative impact of this disease on the body's joints and soft tissues. It is believed that the disease's degenerative effect on articular cartilage is due to a complex combination of both mechanical and biological factors. The lab is working to identify and understand these factors on several length scales. In this particular study, we consider two mathematical models that are relevant to current experimental studies of articular cartilage that aim to understand: (i) the effect of ECM anisotropy on molecular diffusion, and (ii) the effect of pericellular matrix permeability on mechanotransduction in the chondron.

4.2 2-D Models of Local Diffusion in the FRAP Experiment: Circular Bleaching

We consider a 2-D model of fluorescence recovery after photobleaching (FRAP), an experiment that is used to determine local effective diffusion coefficients in soft tissues. Our model incorporates the effect of diffusion anisotropy which is believed to be induced by the ECM anisotropy in articular cartilage. In the photobleaching experiment, fluorescent tracer particles are introduced into a region of tissue and their diffusion is monitored

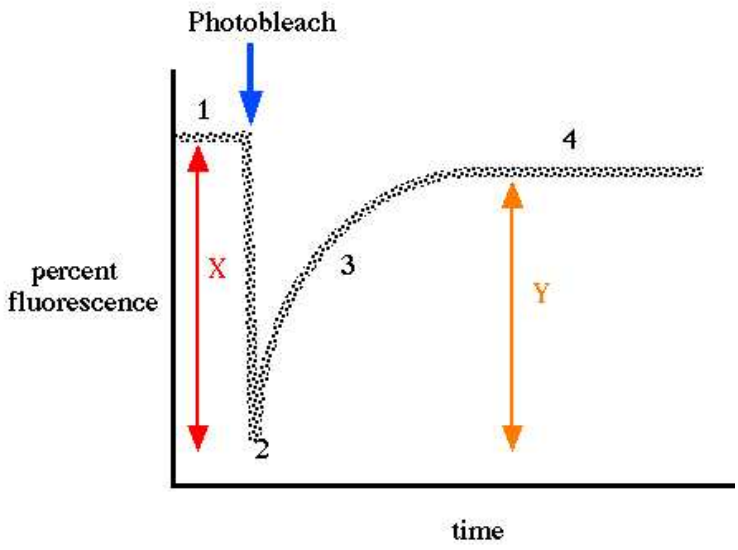


Figure 4.2: An example of a fluorescence output curve in an instantaneous photobleaching experiment.

using confocal microscopy. Knowledge of local diffusion coefficients can be used to quantify the variation of diffusion properties with site in a sample and across sample populations. Since cartilage is aneural and avascular, local diffusion properties impact the transport of nutrients and pharmaceutical agents through the ECM to the cells.

As an extension of previous FRAP models, which considered bleaching of a rectangular region, we consider the case of a circular photobleaching region. The primary advantage of the circular models is that we expect to see an elliptical diffusion profile that, automatically, aligns its long axis to the direction of preferential diffusion.

4.2.1 Mathematical Model - Governing Equations

The mathematical model is formulated based on the Conservation of Mass Balance Equation:

$$\frac{d}{dt}\int_{\Omega} C(\mathbf{x},t)dV = \int_{\Omega} f(\mathbf{x},t)dV - \int_{\partial\Omega} \mathbf{J}\cdot\mathbf{n}dA, \tag{4.1}$$

where $\Omega \subset \mathbb{R}^2$ is a smoothly bounded region representing a sample of cartilage ECM, $C(\mathbf{x},t)$ denotes the concentration (fluorescence intensity) at the point $\mathbf{x} = (x_1, x_2) \in \Omega$, $\mathbf{J}(\mathbf{x},t)$ denotes the flux across the boundary $\partial\Omega$ and $f(\mathbf{x},t)$ is the distributed reaction. Assuming that C and $\mathbf{J}$ are differentiable, we obtain the partial differential equation:

$$\frac{\partial}{\partial t}C(\mathbf{x},t) + \nabla\cdot\mathbf{J}(\mathbf{x},t) = f(\mathbf{x},t) \quad \text{for} \quad \mathbf{x}\in\Omega. \tag{4.2}$$

We assume that the diffusive flux is governed by Fick's Law:

$$\mathbf{J} = -\kappa(\mathbf{x})\cdot\nabla C(\mathbf{x},t), \tag{4.3}$$

where $\kappa(\mathbf{x})$ is the diffusion coefficient.

We model the FRAP photobleaching experiment for cartilage for length scales on which the tissue is assumed to be homogeneous. Consequently, the effective diffusion coefficient is assumed to be constant. In the FRAP experiment, a small area of tissue is exposed to an intense beam of light from a laser microscope causing irreversible photobleaching of the flourophore in that region. An attenuated laser beam is then used

to measure the recovery of the fluorescence in the bleached area due to diffusion of fluorescent molecules from the surrounding unbleached areas.

4.2.2 Instantaneous Bleaching Model

The data from a FRAP experiment in which the photobleaching is applied in a short instant of time is shown in Fig. 4.2. The label **1** indicates the fluorescence level in the sample before the experiment, **2** indicates almost instantaneous photobleaching of a small region of the sample, **3** indicates the percentage of fluorescence recovered over time, and **4** shows the final percentage of fluorescence recovered. The effective diffusion coefficient of the fluorescent tissue can then be computed by fitting a model to the recovery curve for diffusion of fluorescence into the bleached area.

We consider the isotropic 2-D diffusion equation:

$$C_t = \kappa(C_{xx} + C_{yy}) \text{ on } \Omega = [0, R] \times [0, R], t > 0 \tag{4.4}$$

with the homogeneous Dirichlet boundary condition:

$$C(x, y, t) = 0 \text{ for } (x, y) \in \partial\Omega.$$

We model photobleaching of a circle $B_a\left(\frac{R}{2}, \frac{R}{2}\right)$ with radius a in a short instant of time via the initial condition:

$$C(x, y, 0) = \begin{cases} C_0 & \text{for } (x, y) \in B_a\left(\frac{R}{2}, \frac{R}{2}\right) \\ 0 & \text{otherwise} \end{cases} . \tag{4.5}$$

where C_0 is the intensity in the "burned out" region. Via separation of variables, the exact solution of (4.4) is given by:

$$C(x, y, t) = \sum_{n=1}^{\infty} \sum_{n=1}^{\infty} A_{mn} \sin\left(\frac{n\pi}{R}x\right) \sin\left(\frac{m\pi}{R}y\right) \exp\left(-\lambda_{mn}^2 t\right),$$

where:

$$\begin{aligned} \lambda_{mn} &= \pi\sqrt{\kappa\left(\frac{m^2}{R^2} + \frac{n^2}{R^2}\right)} \\ A_{mn} &= \frac{4}{R^2} \int_0^R \int_0^R C(x, y, 0) \sin\left(\frac{m\pi}{R}x\right) \sin\left(\frac{n\pi}{R}y\right) \mathrm{d}y\mathrm{d}x. \end{aligned} \tag{4.6}$$

Inserting the initial condition (4.5) into (4.6) and transforming to polar coordinates (ρ, θ) in the integral, we obtain:

$$\begin{aligned} A_{mn} &= \frac{4C_0}{R^2} \int_0^{2\pi} \int_0^a \rho \sin\left(\frac{\pi m}{R}\left(\rho\cos\phi + \frac{R}{2}\right)\right) \sin\left(\frac{n\pi}{R}\left(\rho\sin\phi + \frac{R}{2}\right)\right) \mathrm{d}\rho\mathrm{d}\phi \\ &= \frac{2C_0}{R^2} \int_0^{2\pi} \int_0^a \rho \left\{\cos\left((m-n)\frac{\pi}{2} + \frac{\rho}{R}W(2\pi - \phi)\right) - \cos\left((m+n)\frac{\pi}{2} + \frac{\rho}{R}W(\phi)\right)\right\} \mathrm{d}\rho\mathrm{d}\phi, \end{aligned} \tag{4.7}$$

where $W(\phi) \equiv \pi(m\cos\phi + n\sin\phi)$.

The mathematical form of (4.7) suggests that the coefficients can be separated into two separate cases:

(i) If n is even then $(m + n) - (m - n) \equiv 2n \equiv 0 \pmod 4$. In this case, it can be shown that:

$$A_{mn} = 0. \tag{4.8}$$

(ii) If n is odd then $(m+n)-(m-n) \equiv 2n \equiv 2 \pmod 4$. In this case:

$$A_{mn} = 4C_0 \int_0^{2\pi} \frac{1}{W^2(\phi)} \left[\cos\left(\frac{(m+n)\pi}{2}\right) - \cos\left(\frac{(m+n)\pi}{2} + \frac{a}{R}W(\phi)\right) - \frac{a}{R}W(\phi)\sin\left(\frac{(m+n)\pi}{2} + \frac{a}{R}W(\phi)\right)\right] \mathrm{d}\phi. \tag{4.9}$$

In regions of articular cartilage near the surface and bone, the collagen fibers tend to have a preferred alignment in directions that are tangential and normal to the interfaces, respectively. As a first approximation, we hypothesize that this alignment will cause the tissue to exhibit anisotropic diffusion with the diffusion coefficient of the form:

$$\kappa(\mathbf{x}) = \begin{bmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{bmatrix} \quad \boldsymbol{\kappa_1, \kappa_2} \ \text{constant} \tag{4.10}$$

Hence, for the anisotropic case, we consider the following 2-D diffusion equation:

$$C_t = \kappa_1 C_{xx} + \kappa_2 C_{yy} \text{ on } \Omega = [0,R]\times[0,R], t > 0 \tag{4.11}$$

with the same initial and boundary condition as in the isotropic case.

Via the coordinate transformation

$$\bar{x} \equiv \frac{x}{\sqrt{\kappa_1}}, \quad \bar{y} \equiv \frac{y}{\sqrt{\kappa_2}}, \quad \bar{C}(\bar{x},\bar{y},t) \equiv C(\sqrt{\kappa_1}\bar{x}, \sqrt{\kappa_2}\bar{y}, t),$$

equation (4.11) can be written as $\bar{C}_t = \bar{C}_{\bar{x}\bar{x}} + \bar{C}_{\bar{y}\bar{y}}$ and the isotropic solution can be employed to obtain:

$$C(x,y,t) = \sum_{n=1}^{\infty}\sum_{n=1}^{\infty} A_{mn} \sin\left(\frac{n\pi}{R}x\right)\sin\left(\frac{m\pi}{R}y\right)\exp\left(-\bar{\lambda}_{mn}^2 t\right), \tag{4.12}$$

with

$$\bar{\lambda}_{mn} = \pi\sqrt{\kappa_1\frac{m^2}{R^2} + \kappa_2\frac{n^2}{R^2}}$$

where the coefficients A_{mn} are given by (4.8) and (4.9).

Since we have a rapidly decaying series in t, we truncated the series (4.12) ($m \le 20, n \le 20$) and evaluated the integral in (4.9) numerically using the trapezoidal rule with $2(m+n)$ points. As we are interested in analyzing the extent to which ECM anisotropy induces preferential molecular diffusion in cartilage, we generated results for the case of relatively weak anisotropic diffusion (Fig. 4.3). We observe that the primary advantage of the circular model over previous rectangular models of instantaneous bleaching as being that the long axis of the elliptical diffusion pattern aligns along the axis of preferential diffusion.

4.2.3 Continuous Bleaching Model

While the circular model of instantaneous bleaching improves upon previous models, the process of instantaneous bleaching has the disadvantage that diffusion recovery occurs on a relatively fast time scale. As such, images in the photobleaching experiment become rather diffuse at very short times making it difficult to detect diffusive anisotropy via image analysis.

Consequently, we consider a model for a continuous photobleaching experiment in which the laser bleaches a very small region of tissue for a continuous period of time. We model this experiment via a point-source bleaching term in the diffusion equation written on the infinite plane. The assumption of an infinite domain is reasonable as the bleached area is much smaller than the area of the tissue sample. The nonhomogeneous isotropic diffusion equation is written in the plane as:

$$C_t = \kappa(C_{xx} + C_{yy}) + q(x,y,t) \text{ on } \mathbb{R}^2, t > 0 \tag{4.13}$$

with homogeneous initial condition $C(x,y,0) = 0$ on $\mathbb{R}^2$.

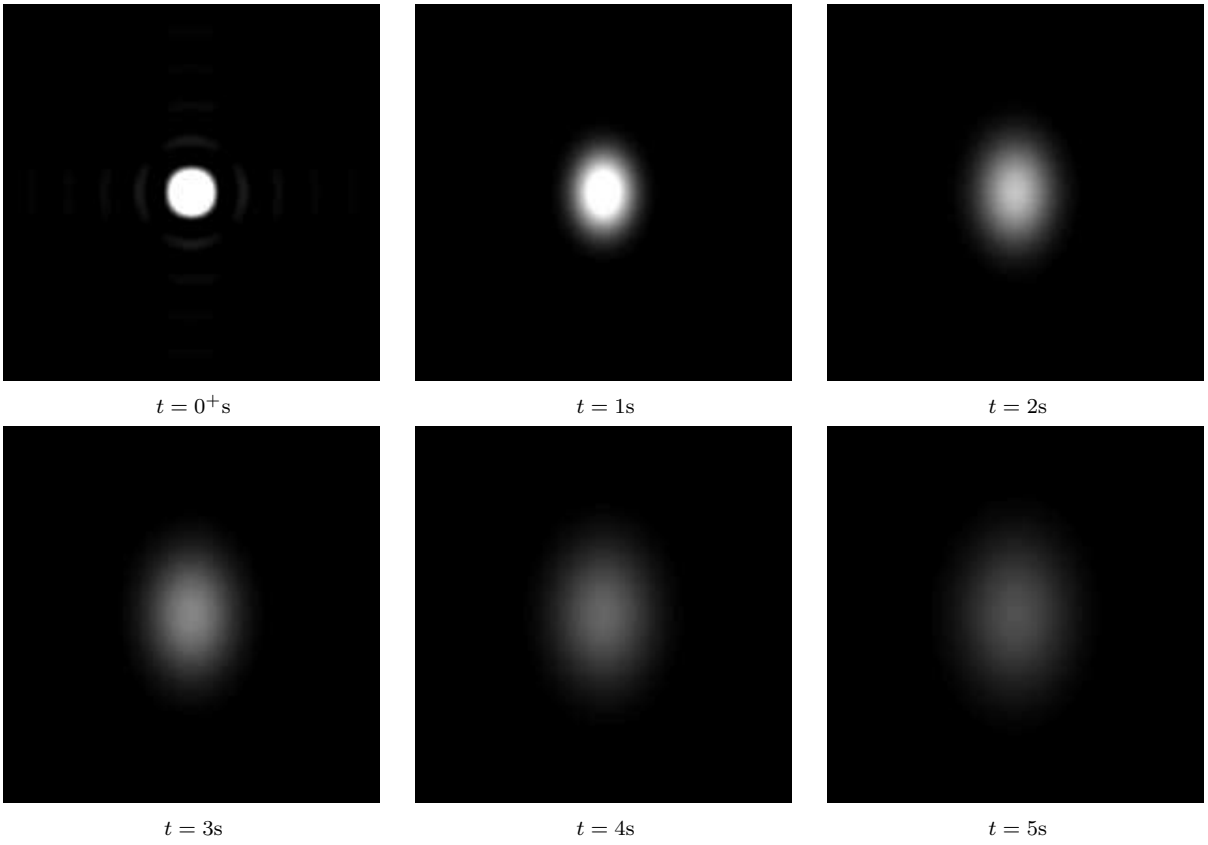


Figure 4.3: Simulations of instantaneous photobleaching images for a square tissue sample of size $R = 100\mu$m with a circular photobleaching region of radius $a = 5\mu$m. The diffusion coefficients where taken as $\kappa_1 = 10\mu\text{m}^2 s^{-1}$ and $\kappa_2 = 20\mu\text{m}^2 s^{-1}$ and the intensity in the bleached region was set at $C_0 = 255$

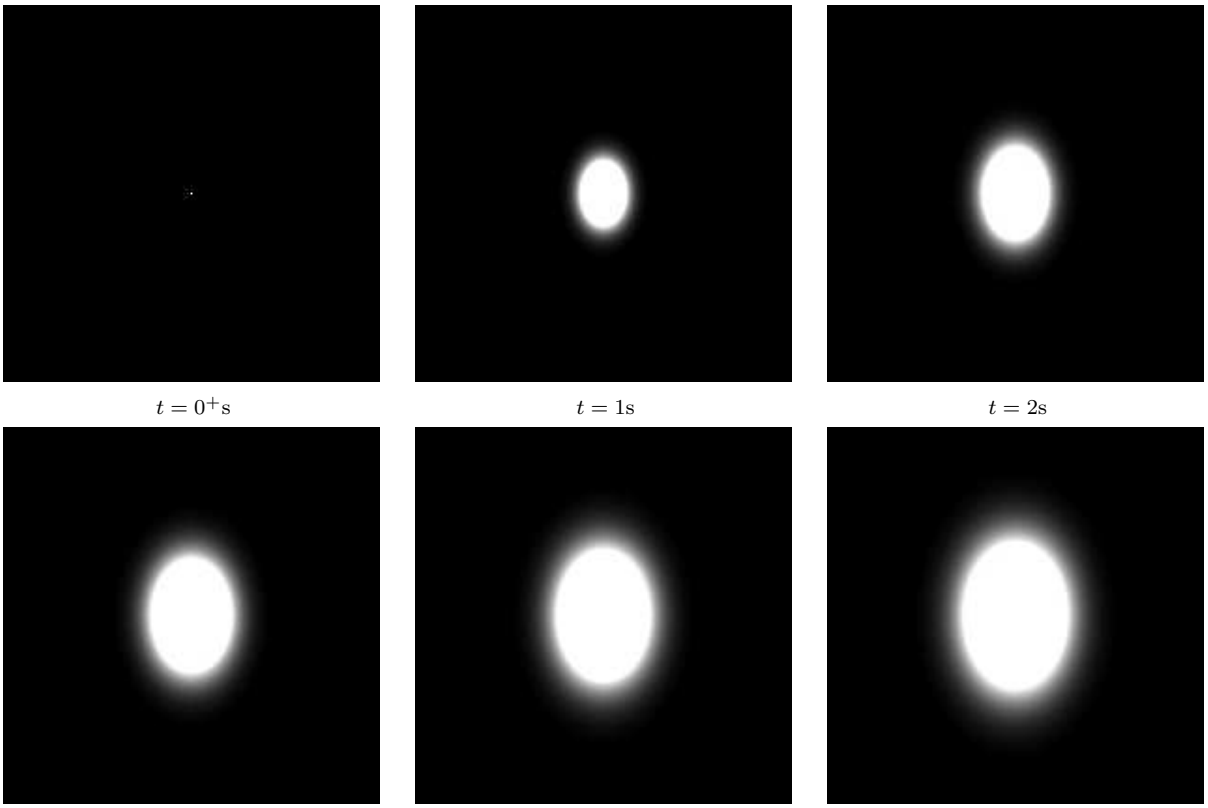


Figure 4.4: Simulations of continuous photobleaching images for an infinite sample with a continuous bleaching of power $Q_0 = 50,000$ at the origin. The diffusion coefficients where taken as $\kappa_1 = 10\mu\text{m}^2 s^{-1}$ and $\kappa_2 = 20\mu\text{m}^2 s^{-1}$ and images are shown on the domain $[-50\mu\text{m}, 50\mu\text{m}] \times [-50\mu\text{m}, 50\mu\text{m}]$

The exact solution of (4.13) can be written as

$$C(x,y,t) = \int_0^t \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S_2(x-\xi, y-\eta, t-s) q(\xi,\eta,s) \mathrm{d}\xi \mathrm{d}\eta \mathrm{d}s, \tag{4.14}$$

where S_2 is the heat kernel

$$S_2(x,y,t) = \frac{1}{4\pi k t} \exp\left(\frac{x^2+y^2}{4kt}\right).$$

To model continuous bleaching at a point, we let $q(x,y,t) \equiv Q_0 \delta(x)\delta(y)$ which represents a source of strength Q_0 located at the origin, where Q_0 is the bleaching power of the laser. For this model, (4.14) reduces to:

$$C(x,y,t) = Q_0 \int_0^t \frac{1}{4\pi k(t-s)} \exp\left(-\frac{x^2+y^2}{4k(t-s)}\right) \mathrm{d}s = \frac{Q_0}{4\pi k} \mathrm{E}_1\left(\frac{x^2+y^2}{4tk}\right),$$

where E_1 is the exponential integral function:

$$\mathrm{E}_1(x) \equiv \int_1^{\infty} \frac{e^{-xt}}{t} \mathrm{d}t = \int_x^{\infty} \frac{e^{-t}}{t} \mathrm{d}t$$

For the anisotropic case, we have the differential equation for fluorescence intensity in the anisotropic infinite area:

$$C_t = \kappa_1 C_{xx} + \kappa_2 C_{yy} + q(x,y,t) \text{ on } \mathbb{R}^2, t > 0 \tag{4.15}$$

with the same initial condition and source term as in the isotropic case. Via the coordinate transformation:

$$\bar{x} \equiv \frac{x}{\sqrt{\kappa_1}}, \quad \bar{y} \equiv \frac{y}{\sqrt{\kappa_2}}, \quad \bar{C}(\bar{x},\bar{y},t) \equiv C(\sqrt{\kappa_1}\bar{x}, \sqrt{\kappa_2}\bar{y}, t)$$

equation (4.15) becomes:

$$\bar{C}_t = \bar{C}_{\bar{x}\bar{x}} + \bar{C}_{\bar{y}\bar{y}} + \frac{1}{\sqrt{\kappa_1\kappa_2}} q,$$

which is formally equivalent to (4.13). Hence the anisotropic solution is given by:

$$C(x,y,t) = \frac{Q_0}{4\pi\sqrt{\kappa_1\kappa_2}} \mathrm{E}_1\left(\frac{x^2/\kappa_1 + y^2/\kappa_2}{4t}\right). \tag{4.16}$$

Equation (4.16) can be evaluated directly using the `expint` command in MATLAB. Typical results are shown in Fig. 4.4. We observe that the continuous bleaching source term in the model results in much sharper images as compared to the instantaneous bleaching model. The bleached region grows as an ellipse and the rate of expansion of the ellipse can be determined from the level curves of the solution (4.16), given by:

$$\frac{x^2/\kappa_1 + y^2/\kappa_2}{4t} = \text{const.}$$

The time scale on which the ellipse grows is much slower than the recovery time scale in the instantaneous bleaching model (Fig. 4.3). Consequently, the continuous bleaching model shows promise for improving the quantification of diffusion anisotropy in the FRAP photobleaching experiment via modification of the bleaching protocol to one in which the laser bleaches a smaller region for a continuous period of time.

4.2.4 Reaction Model of Continuous Bleaching

We briefly considered a third model of continuous photobleaching, based on [1], by replacing q in (4.15) with a reaction term that is proportional to C. Such a model may be more realistic in situations where there is a limited supply of fluorescent molecules, as in the case of bleaching a finite region with an impermeable boundary like an individual cell.

Let $\Omega \equiv [-\frac{R}{2}, \frac{R}{2}] \times [-\frac{R}{2}, \frac{R}{2}]$ and let B_a be a circle of radius a centered at the origin. We model continuous bleaching as a process by which the laser is continuously burning out a certain percentage of glowing molecules in B_a. The anisotropic diffusion equation is:

$$C_t = \kappa_1 C_{xx} + \kappa_2 C_{yy} + gC \text{ on } \Omega, t > 0, \quad C(x, y, 0) = C_0 \text{ on } \Omega, \tag{4.17}$$

with:

$$g(x, y, t) = \begin{cases} -g_0 & \text{in } B_a \\ 0 & \text{otherwise} \end{cases}$$

Equation (4.17) was solved numerically using a finite difference method. We employed central differences in space and a Crank-Nicolson scheme in time. Typical results are shown in Fig. 4.5. We observe that the photobleaching images remain relatively static in time as the transient changes in concentration are localized to the bleached circle. As such, quantification of diffusion coefficients via image analysis may prove cumbersome if this model is realistic. However, we believe that in FRAP experiments of the ECM, there are a large number of molecules available to make the continuous bleaching model (Fig. 4.4) a more accurate representation of the process of continuous bleaching. Future experiments in the Orthopaedic Research Lab will be designed to incorporate continuous bleaching and refine the models presented in this study.

4.3 1-D Spherical Model for Mechanotransduction in a Chondron

We now consider a model for mechanical signal transmission in a chondron. The chondron consists of a cell (chondrocyte) that is encapsulated by a pericellular matrix (PCM). It is believed that the functional role of the PCM is to protect the cell and allow mechanical signals to reach the cell from the ECM. Some distinct properties of the PCM are that it is much stiffer than the cell and, in contrast to the ECM, is dominated by type-VI collagen.

4.3.1 1-D Spherical Model of the Chondron

In our model, we assume that a time-varying sinusoidal signal has been transmitted throughout the ECM and arrived at the chondron. We model the chondron as a spherical cell with an attached layer that represents the PCM and introduce spherical coordinates (ρ, θ, ϕ). We model the cell as a linear and isotropic biphasic continuum and assume that all deformation and fluid flow occurs in the radial direction ρ (Fig. 4.6). Under these assumptions, the governing equations of linear biphasic theory can be reduced to:

$$\partial_t u = \begin{cases} k_C H_C^A \left(\rho^{-2}\partial_\rho(\rho^2 \partial_\rho u) - 2\rho^{-2} u\right) & 0 < \rho < a \\ k_P H_P^A \left(\rho^{-2}\partial_\rho(\rho^2 \partial_\rho u) - 2\rho^{-2} u\right) & a < \rho < b \end{cases}, \quad t > 0 \tag{4.18}$$

$$p = \begin{cases} H_C^A(2\rho^{-1}u + \partial_\rho u) + f_C(t) & 0 < \rho < a \\ H_P^A(2\rho^{-1}u + \partial_\rho u) + f_P(t) & a < \rho < b \end{cases}, \quad t > 0 \tag{4.19}$$

where the unknowns are the displacement u and pore pressure p. The subscripts C and P denote quantities associated with the cell $(0 < \rho < a)$ and PCM $(a < \rho < b)$, respectively and $f_C(t)$ and $f_P(t)$ are arbitrary functions of time. We model the arrival of a mechanical signal at the chondron via a boundary condition at $\rho = b$ with sinusoidal input:

$$u(b, t) = u_0 \sin(\omega t) \equiv I(t) \qquad t > 0 \tag{4.20}$$

Along the cell-PCM interface $(\rho = a)$, the biphasic jump conditions for the solid phase reduce to:

$$u(a^+, t) = u(a^-, t) \tag{4.21}$$

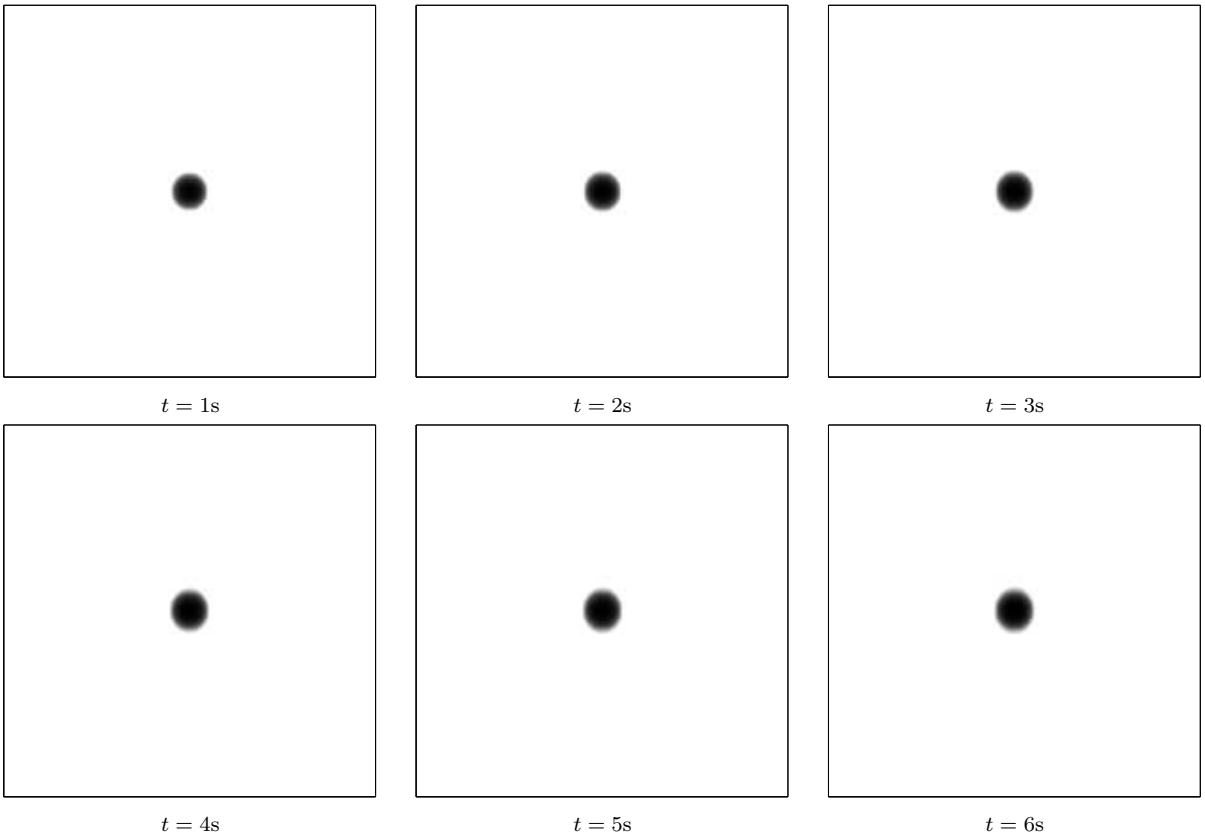


Figure 4.5: Simulations of images for a reaction model of continuous photobleaching for square sample of width $R = 100\mu$m and radius of the bleached circle of $a = 5\mu$m. The diffusion coefficients were taken as $\kappa_1 = 10\mu\text{m}^2 s^{-1}$ and $\kappa_2 = 20\mu\text{m}^2 s^{-1}$ and the bleaching intensity and decay coefficient were $C_0 = 255$ and $g_0 = 16\text{s}^{-1}$, respectively. The parameters in the finite difference scheme were $\Delta t = \frac{1}{30}s$ and $\Delta x = \Delta y = 1\mu$m.

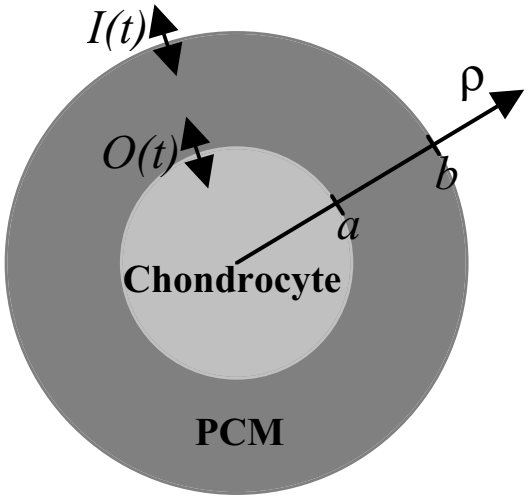


Figure 4.6: A spherical model of the chondron in articular cartilage. The chondron is modeled as a biphasic spherical cell (chondrocyte) with attached biphasic spherical layer. A displacement input signal is applied at the outer boundary ($\rho = b$) and signal transmission is measured by calculating the transmitted output signal at the cell-PCM interface ($\rho = a$).

$$H_P^A \partial_\rho u(a^+,t) - H_C^A \partial_\rho u(a^-,t) = \frac{2\lambda_C}{a} u(a^-,t) - \frac{2\lambda_P}{a} u(a^+,t) \tag{4.22}$$

The biphasic jump condition for continuous pressure across the interface can be enforced through an appropriate choice of the arbitrary functions $f_C(t)$ and $f_P(t)$ in (4.19).

For this model, the displacement $u(\rho,t)$ and pressure $p(\rho,t)$ uncouple and our signal transmission model consists of the solution of (4.18) subject to (4.20)-(4.22). In particular, we are interested in the relation between the input signal in (4.20) and the (output) displacement signal at the cell-PCM interface:

$$O(t) \equiv u(a,t) = B\sin(\omega t + \delta) \tag{4.23}$$

We perform a parametric analysis on the amplitude B of the transmitted signal. This amplitude measures the performance of the PCM as a signal transmitter as a function of the forcing frequency and material parameters in the model.

4.3.2 Finite Difference Solution

The spherical chondron model is a linear interface problem and, under certain circumstances, has a solution representation in terms of an eigenfunction series expansion. However, the condition for real eigenvalues in the series expansion (self-adjointness of the operator) is that $k_C = k_P$. Using our numerical model, we are able to relax this condition. Hence, the focus of this study will be to quantify the effect of permeability on signal transmission in the PCM.

Equation (4.18) is re-written as:

$$\partial_t u = \alpha\left(\partial_\rho^2 u + \frac{2}{\rho}\partial_\rho u - \frac{2}{\rho^2}u\right) \qquad 0<\rho<b, \quad \text{where: } \alpha = \begin{cases} k_C H_C^A & 0<\rho<a \\ k_P H_P^A & a<\rho<b \end{cases} \tag{4.24}$$

We employ a forward finite-difference approximation for the time derivative:

$$\partial_t u(\rho,t+\Delta t) = \frac{u(\rho,t+\Delta t) - u(\rho,t)}{\Delta t} \tag{4.25}$$

and a centered finite-difference approximation for the spatial derivatives:

$$\partial_\rho^2 u(\rho,t+\Delta t) = \frac{u(\rho+\Delta\rho,t+\Delta t) - 2u(\rho,t+\Delta t) + u(\rho-\Delta\rho,t+\Delta t)}{(\Delta\rho)^2} \tag{4.26}$$

$$\partial_\rho u(\rho,t+\Delta t) = \frac{u(\rho+\Delta\rho,t+\Delta t) - u(\rho-\Delta\rho,t+\Delta t)}{2\Delta\rho} \tag{4.27}$$

We also assume that $u(0,t)=0$ since there is no displacement at the center of cell ($\rho=0$).

Introducing a regular mesh in time and space, let $u_i^j = u(\rho_i,t_j)$ where $\rho_i = ih$, $t_j = jk$, $h=\Delta\rho$, and $k=\Delta t$. Equation (4.24) can be written as:

$$\frac{u_i^{j+1}-u_i^j}{k} = \alpha\left[\frac{u_{i+1}^{j+1} - 2u_i^{j+1} + u_{i-1}^{j+1}}{h^2} + \frac{2}{\rho_i}\frac{u_{i+1}^{j+1} - u_{i-1}^{j+1}}{2h} - \frac{2}{(\rho_i)^2}u_i^{j+1}\right] \tag{4.28}$$

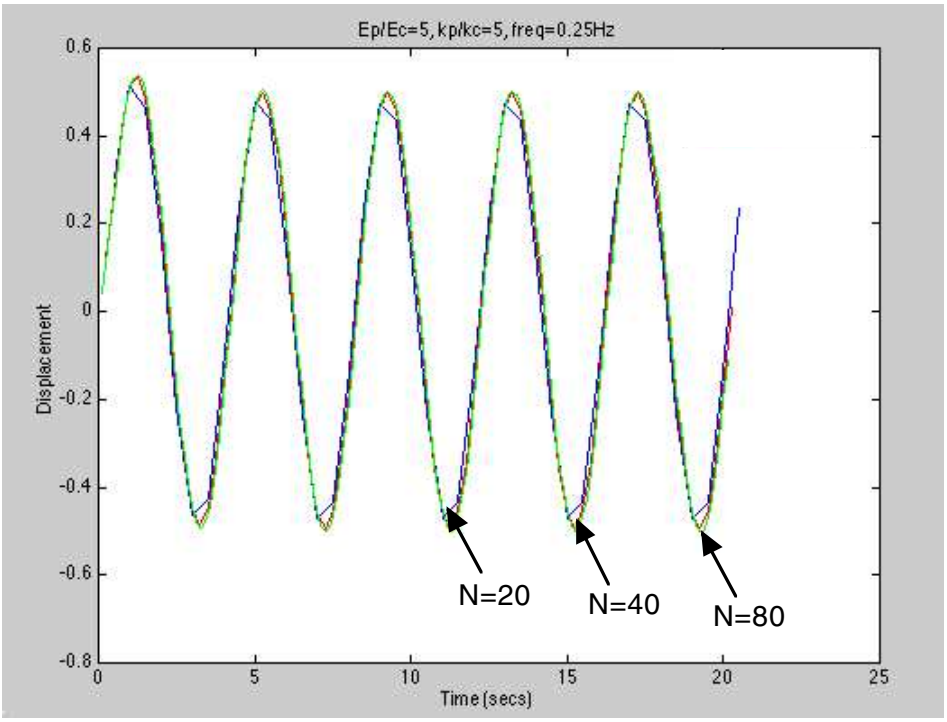
To enforce the interface conditions at $\rho=a$, we mesh so that no point on the mesh is coincident with the interface. Denoting the index i of the mesh point immediately to the left of the interface by A, we discretize the interface condition in (4.22) as:

$$H_P^A\left[\frac{u_{A+2}^{j+1} - u_{A+1}^{j+1}}{h}\right] - H_C^A\left[\frac{u_A^{j+1} - u_{A-1}^{j+1}}{h}\right] = \frac{2\lambda_C}{a}u_A^{j+1} - \frac{2\lambda_P}{a}u_{A+1}^{j+1} \tag{4.29}$$

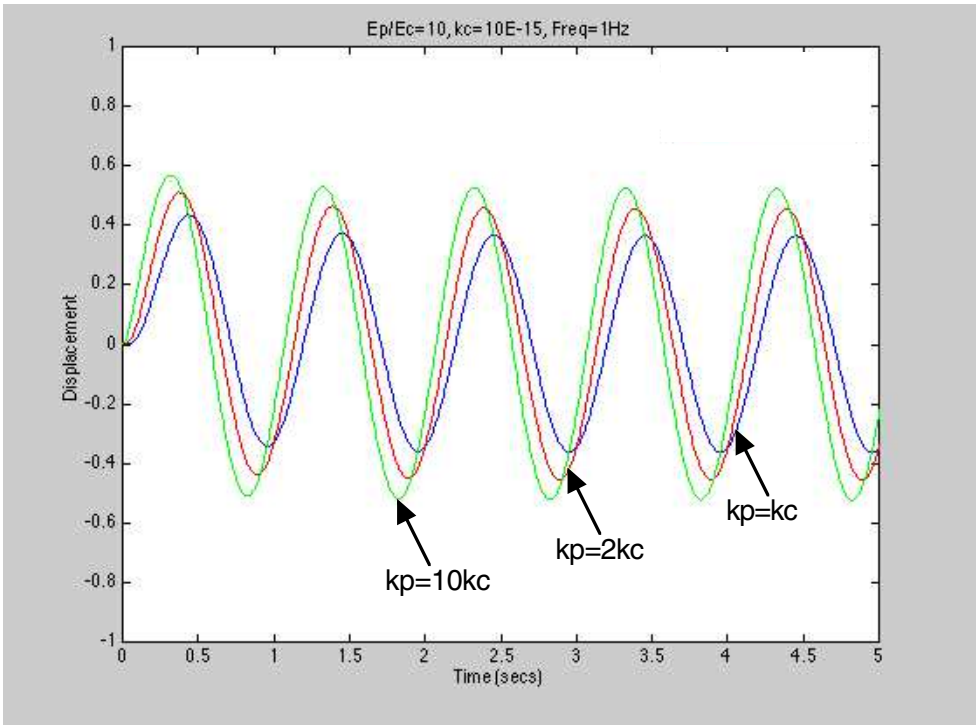
and also enforce the interface condition in (4.21):

$$u_{A+1}^{j+1} = u_A^{j+1} \tag{4.30}$$

Hence, our scheme consists of solving (4.28) with $\alpha = k_C H_C^A$ for $i=1,\cdots,A-1$, equation (4.29) for $i=A-1$, equation (4.30) for $i=A$ and equation (4.28) with $\alpha = k_P H_P^A$ for $i=A+1,\cdots,N-1$. At $i=1$ ($\rho=0$) and $i=N$ ($\rho=N$) the appropriate boundary conditions are enforced. Marching in time, we then solve a linear algebraic system at each time step using MATLAB to obtain the numerical solution of the spherical chondron model for a specific set of parameter values.



(a)



(b)

Figure 4.7: Finite difference simulations of transmitted displacement signals at the cell-PCM interface ($\rho = a$): (a) Numerical convergence of the scheme is rapid and a steady state signal is achieved within a short period of time. (b) The effect of increasing permeability of the PCM on the transmitted signal at the cell-PCM interface.

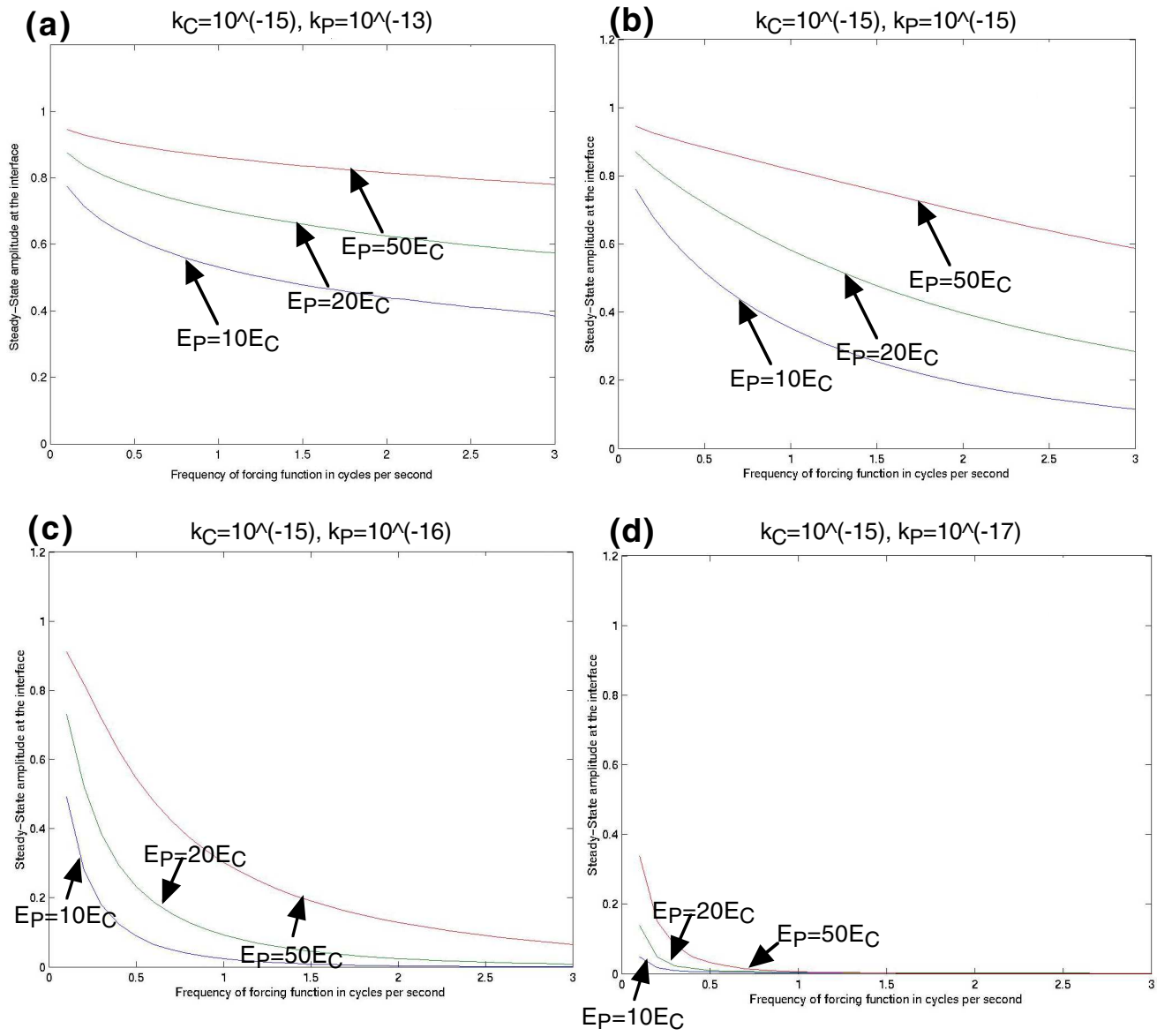


Figure 4.8: A parametric analysis of the effect of permeability on the amplitude of the transmitted signal at the cell-PCM interface. The amplitude of the steady-state transmitted displacement signal at the cell-PCM interface is plotted as a function of the frequency of the applied displacement signal at the outer PCM boundary. Each point on a graph represents one run of the finite difference code. (a) $k_P = 100k_C$ (b) $k_P = k_C$ (c) $k_P = 0.1k_C$ (d) $k_P = 0.01k_C$

4.3.3 Results: Parametric Analysis of the Effect of Permeability

Since the PCM serves as a protective layer for the cell, its elastic modulus is significantly stiffer than that of the cell. Initial micropipette aspiration experiments at the Orthopaedic Research Lab have indicated that the ratio of Young's moduli E_P/E_C can be as large as 100 in healthy tissue, and somewhat diminished in the case of osteoarthritis. We use our finite difference model to simulate signal transmission in a chondron with the representative properties $a = 10\mu$m, $b = 12.5\mu$m, $\nu_C = 0.45$, $\nu_P = 0.1$ for three different moduli ratios $E_P/E_C = 10, 20, 50$. We consider the effect of changing the permeability of the PCM k_P relative to the permeability of the cell k_C in a range of frequencies that is representative of human motion (f = 0-3Hz). To quantify signal transmission, the input amplitude was taken as $u_0 = (b-a)/10$. For a typical case, we found that our numerical scheme converged rapidly (Fig. 4.7a) and $N = 80$ was sufficient for our parametric analysis. We see that the displacement signal at the cell-PCM interface has a transient component that rapidly tends the the steady-state oscillatory signal. The effect of increasing the permeability of the PCM is demonstrated in Fig.4.7b. We observe that increased permeability of the PCM enhances signal transmission through the PCM to the cell.

For a comprehensive analysis of the effect of permeability on signal transmission, we ran our code for 30 values of forcing frequency in the range $0-3$Hz for three different Young's moduli ratios and several permeabilities. In the case of equal permeability, amplitudes were compared to an analytical series solution based on an eigenfunction expansion and found to agree. The effect of permeability on the amplitude B of the steady state transmitted signal at the cell-PCM interface is shown in Fig. 4.8. We observe that decreasing the permeability of the PCM relative to the cell by two orders of magnitude has a significant detrimental effect on signal transmission in the chondron (Fig. 4.8c-d). As the PCM permeability increases up to 100 times the cell permeability, signal transmission is enhanced. In our simulations the optimal case, with $80-100\%$ signal transmission, occurs when the PCM permeability is 100 times that of the cell and the PCM is 50 times stiffer than the cell (Fig. 4.8a).

It is interesting to observe that, given recent experiments, $E_P/E_C \approx 50$ is a reasonable representation of the ratio of stiffness moduli for chondrons from healthy cartilage. Consequently, our spherical chondron model parametric analysis indicates that a relatively large ratio of PCM to cell permeability enhances signal transmission in the chondron. While steady state permeation is typically used to determine the permeability of the ECM in cartilage, this technique is limited to the length scale of a layer of cartilage. In future work, the Orthopaedic Research Lab will attempt to design microscopic experiments that allow for in vitro determination of permeability in the chondron. For the case of optimal transmission of displacement signals, an analysis of the normal stress at the cell-PCM interface will also be performed to ensure that the cell is not exposed to excessive amounts of normal stress.

Bibliography

[1] Reiner Peters, Axel Brünger, Klaus Schulten *Continuous fluorescence microphotolysis: A sensitive method for study of diffusion processes in single cells,* Proc. Natl. Acad. Sci. USA, Vol. 78, No.2, pp. 962-966, February 1981, Biophysics

Report 5

Recognizing Sand Ripple Patterns from Side-scan Sonar Images

John David[1], Gunay Dogan[2], Chiu Yen Kao[3],
Pakinee Suwannajan[4], Yevgeny Goncharov[5]

Problem Presenter:
Yu Chen
SUMMUS Inc.

Faculty Consultants:
Zhilin Li and Kazi Ito

Abstract

Sand ripples are an important aspect of the seafloor makeup. Being able to detect them from sonar imaging is valuable in understanding the geologic makeup of the seafloor. They have unique properties which differentiate them from other areas of the seafloor. The first property is a dominant direction of the ripple. The second property is a continuity of the ripples. Based on these properties we developed two tests to allow one to mathematically detect these ripples from large sonar images.

5.1 Introduction and Motivation

In the past two years Summus worked on a research program sponsored by the Office of Navel Research to investigate side-scan sonar image processing technologies for the purpose of recovering structure and geological information on the seabed. Such structure and geological information are crucial for US Navy's bottom mapping task. In such task sand ripples are one of most important bottom types.

Recognizing the pattern of these sand ripples from large images is difficult and requires several different methods of analysis. Since scale is important to the recognition of these patterns the first step is to decompose the images using "steerable pyramid" filter[3]. From these decomposed images the "oriented energy" is measured to examine the directionality of the images [1]. This information is represented in a histogram. Our task was to analyze the histograms in order to accurately identify dominant orientation in the scaled images. The

[1]North Carolina State University
[2]University of Maryland, College Park
[3]University of California, Los Angeles
[4]University of Colorado, Boulder
[5]University of Illinois Chicago

images that display a dominant direction in the histogram are then analyzed for spatial coherence in order to detect the continuity of the image, thus giving a way to differentiate between ripple patterns and nonrippled but oriented patterns.

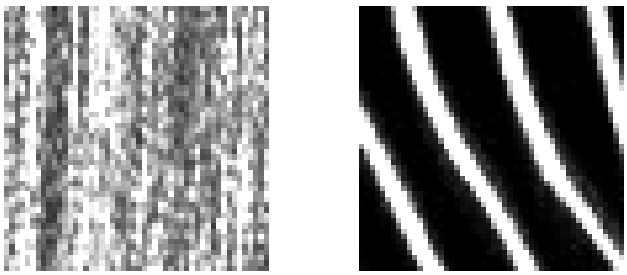


Figure 5.1: Examples of (a)nonsand-ripple image. (b)sand-ripple image

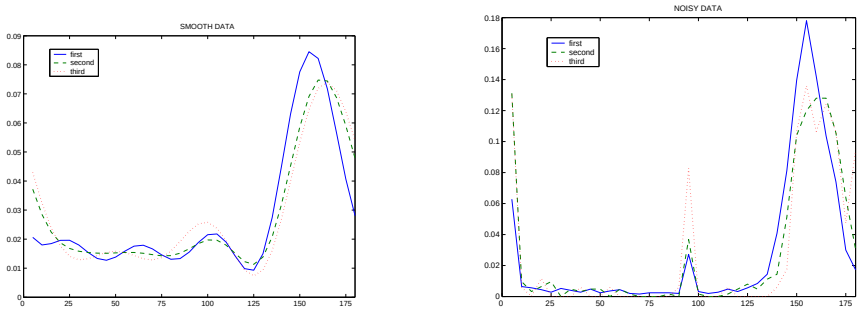


Figure 5.2: (a). A Histogram of the noisy data (original) and (b) the histogram after fast Fourier transform.

5.2 Methodology

5.2.1 Histogram Analysis

Our first task was to analyze the histograms. The histograms represent the data collected from calculating the oriented energy of the image of different scales. For each image there were three levels of decomposition of the data The x-axis represent the degree of orientation and the y-axis represent the energy. For each image there were three levels of decomposition of the data, each with a histogram. Although finding the maximum energy is simple, the question is finding whether this maximum represents a clear orientation of the image. A maximum that represents a clear orientation of the image may be called a dominant peak. First to eliminate some of the noise in the data we performed a fast Fourier transform (fft) on the data. We eliminated the higher frequencies and performed an inverse fast Fourier transform on the data to smooth the curve.

From this we designed two criteria to determine whether the maximum was dominant. First we calculated the height of the maximum relative to the average of the data, H_avg.

$$\text{Relative Height}(Hrel_i) = \frac{H_i - H_{avg}}{H_{avg}} \tag{5.1}$$

This method gives a criteria to determine whether the peak is dominant. From experimentation we found that a peak with a relative height greater than 0.3 was often dominant. However peaks with high relative heights but narrow bases often did not reveal dominant orientation in the original images. So we calculated the ratios of the areas under the peaks,R_i =Area under the peak i /Area under the whole graph.

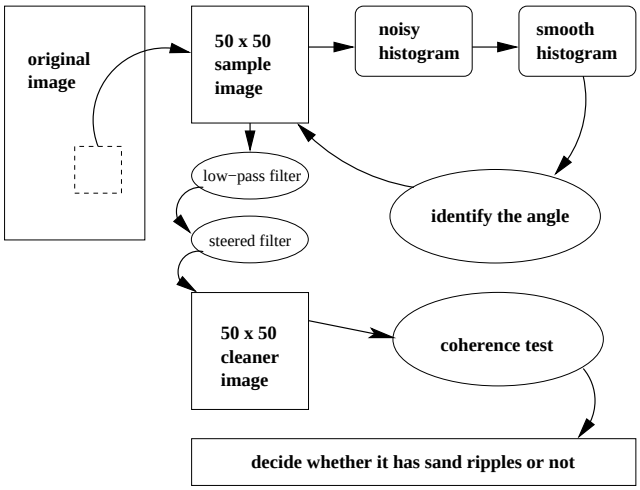


Figure 5.3: Chart shows how the images get determined

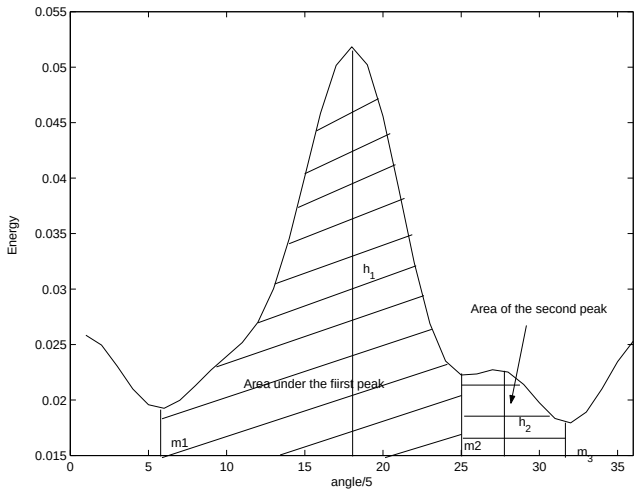


Figure 5.4: Chart describes the histogram

A value of area ratio of peak greater than 0.3 often indicated a dominant peak. Using these criterion we determine whether each histogram has a dominant peak. The final step for analyzing orientation is comparing the dominant peaks of the various levels of decomposition. While the first level had the most data it often also had the most noise. Histograms with dominant peaks at the same orientation at both the first and the second level often had clear orientation in their images. We tested each level data and then compared them to see if the orientations from each level were consistent. A deviation of +5 or -5 degrees of the orientations are accepted. Since level 1 and 2 contain more information than level 3, we will accept the orientation that agrees in both level 1 and 2, no matter what orientation level 3 gives or even when level3 does not detect any orientations.

5.2.2 Result From Histogram Analysis

We tested those criterion to the histograms from various ripple images provided by our presenter. Here are orientations (degree) returned from each level and compared to the original images.

Image's name	Level1	level2	level3	Images has Ripples	Angle agree with image
09mar12xR3500C100ap	90	90	90	Yes	Yes
b2eCutDno	45	30	-	Yes	Yes
bigSandrip	155	160	165	Yes	Yes
106-1024-1Cut	15	15	-	Yes	Yes
spatialCohDno	90	95	95	No	No

The corresponding images and histograms are as below.

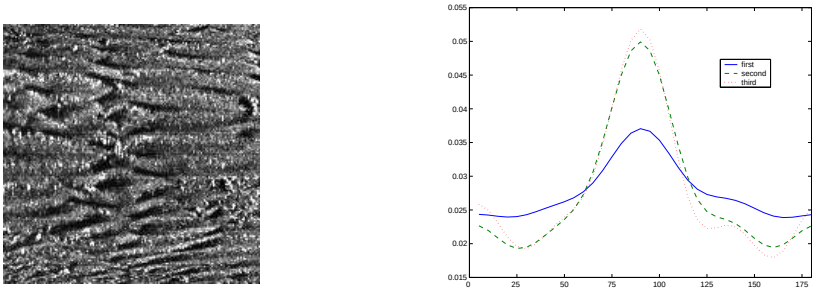


Figure 5.5: 09mar12xR3500C100ap and its histogram.

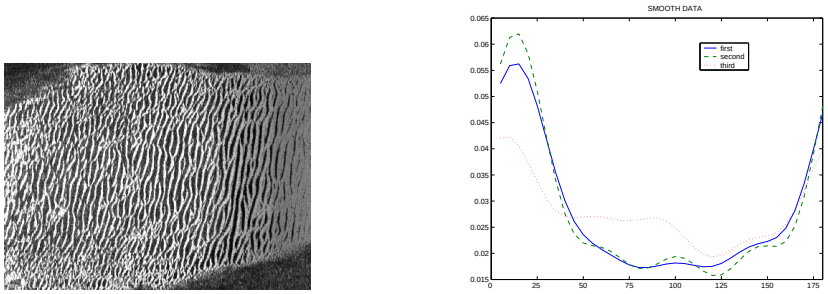


Figure 5.6: 106-1024-1Cut and its histogram .

The results show that these criterion do relatively well in specifying the dominant peak from the histogram. Most of the cases return the same orientation at least from level 1 and 2 and they agree with their images. The question remains what to do with histograms that give different orientations in the different levels. However, for some images for example "bigSandrip", The 155, 160 and 165 degree orientations are returned from level 1,2, and 3 respectively. This image shows clear sand ripples and supports the idea that a differnce of five degrees can be disregarded.

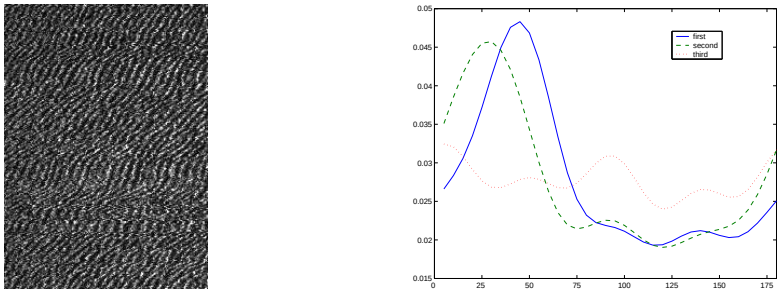


Figure 5.7: b2eCut and its histogram .

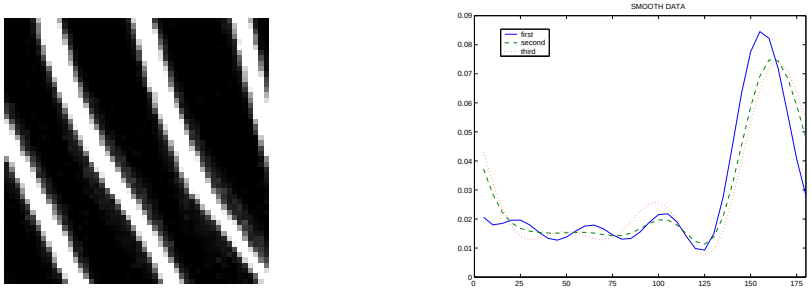


Figure 5.8: bigSandrip and its histogram .

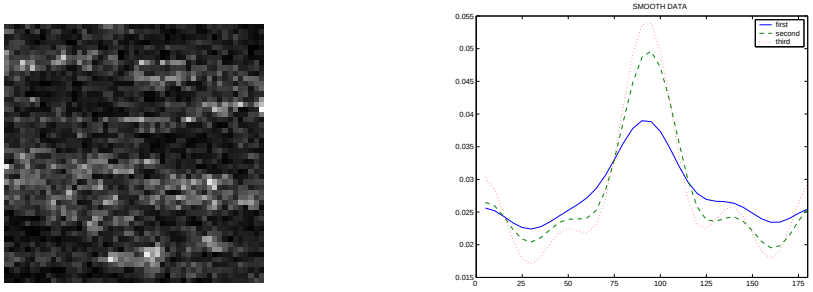


Figure 5.9: SpatialCohDno and its histogram .

5.3 Spatial Coherence Test

Because the presence of dominant orientation of an image does not necessarily means it is an image of a sand ripple, we needed to develop a test to filter out "false alarm" pictures (fig. 9 shows such a picture).

First, we can see that one of the important properties of a sand ripple pattern is a continuity of its structure. Intuitively what this means is ripple patterns look a good deal like continuous functions. Fig. 9, a non-ripple pattern,does not possess such property. First we rotate the image to get a horizontal orientation for all the images. To check this spatial coherence we shift the image along the dominant orientation and evaluate correlation between the original and the shifted image. A picture which possesses spatial coherence will have better correlation with the shifted counterpart than a picture which does not have such property. In our implementation of the procedure an image is rotated first so that it has horizontal orientation. Then we shift it a minimal distance, ie. 1 through 5 pixels. We have evaluated the correlation in a number of experiments and found that threshold of 0.3, for a shift of five pixels works very well to differentiate between such images as fig. 8 and fig 9.

5.3.1 Results From the Coherence Test

Here are the groups of data. First are results from the non-ripple patterns. Their histograms return the degree of orientation but the low coherence number shows that there is no ripple. The second data are from the true ripple sample, the angle and the higher correlation number reveal the correct orientation and the existence of the ripple patterns. The third group did not pass the orientation test, so we did not use the correlation test.

Image's name	Location	angle	correlation number	Ripple?	Agree original image
09mar12xBigCut	(100,200)	95	0.0545	No	Yes
09mar13xCut2	(150,100)	95	0.1225	No	Yes
09mr13xCut2	(100,200)	95	0.0193	No	Yes
106-0920-1False	(200,300)	95	0.1929	No	Yes
09mar12xBigCut	(300 ,100)	85	0.3756	Yes	Yes
106-1024-1Cut	(250,250)	10	0.4042	Yes	Yes
106-1034-1-13000-400	(100,100)	25	0.5384	Yes	Yes
b2-ecut	(50,100)	35	0.5063	Yes	Yes
SandRipplesBig	(50,150)	150	0.627	Yes	Yes
09mar12xBigCut	(450,200)	-	-	No	Yes
09mar13xCut2	(150,100)	-	-	No	Yes
SansRippleBig	(350,350)	-	-	No	Yes

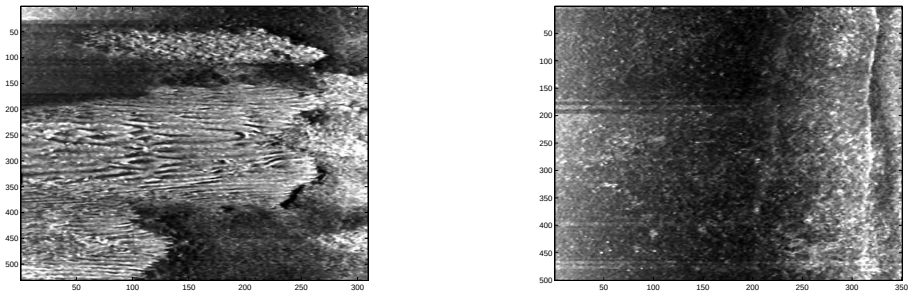


Figure 5.10: 09mar12xBigCut and 09mar13xCut2.

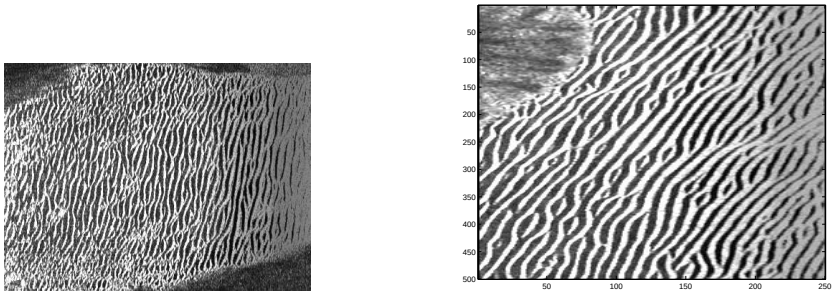


Figure 5.11: 106-1024-1Cut and 106-1034-1-13000-400 .

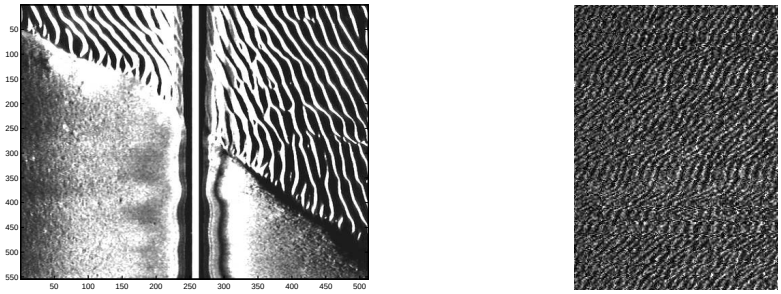


Figure 5.12: SandRipplesBig and b2eCut .

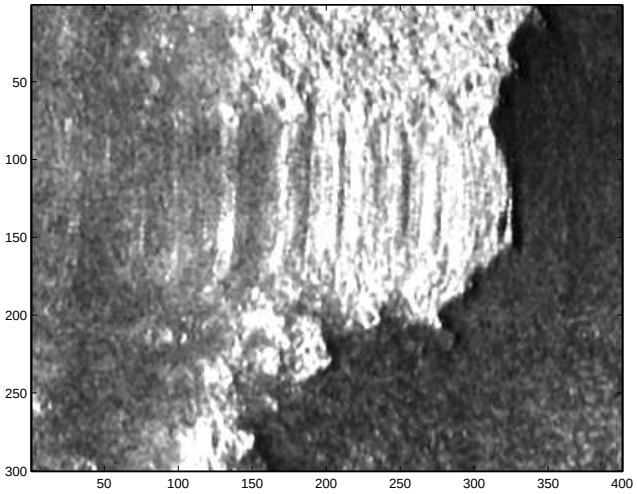


Figure 5.13: Chart describes the histogram

5.4 Future Work

The combination of histogram analysis and spatial coherence anlysis has been shown to work well through the experimentation carried out so far. However, it is still necessary to perform further experimentation to state the limitations of the method more conclusively. The number of test cases corresponding to each category should be increased and the test results should be reviewed carefully. Both levels of analysis mainly reflect a straightforward implementation of the underlying idea. There is still some flexibility to improve the performance of the method via fine tuning of the paramenters and modifying certain components of the algorithm. Extensive testing is certainly necessary to find the right parameters (for example, thresholds). The best choice between using all the angles from the histogram analysis or a weighted average of them or a range including the peak angles may also be explored. The effect of a more accurate itegration scheme for the peak areas may also be checked, as well as the effect of choosing a different displacement between cross sections (i.e. one, two or several number of points).

Second, even though in most cases non-rippled butoriented patterns are successfully filtered out through the spatial coherence test, some may have spatial coherence as strong as sand ripples and therefore can pass the procedure. To deal with this problem we can note that another important property of a sand ripple is its uniform structure, i.e. distance between ripples and the height of the ripples should be approximately the same. Therefore, if an image passes the orientation and coherence tests, we can test the image on "uniformity". This is done by taking several slices and finding variances of distances between maximums of slices and heights of their maximums.

Another possibility to explore is to use machine learning methods to detect sand ripples in the given image segments. Ther are a number of machine learning techniques used for pattern recognition and they have been known to produce good results at certain implications. Neural networks and support vector machines are used extensively for this purpose and it is expected that they may perform well for this problem since the problem is a relatively simple case of pattern recognition. These two methods are also desirable for their robustness properties; they produce good results with noisy data too.

5.5 Conclusion

We have developed two methods to test for sand ripples in sonar images. Based on the fact that the ripple patterns have a dominant orientation and are "quasi-linear" [1], we used histogram analysis to detect possible candidates for ripple patterns. From these candidates we used the fact that ripple patterns are a good deal like continuous functions to test the spatial coherence of the images. From these test we developed a relatively good way to begin to analyze these images for sand ripples.

Acknowledgments

We would like to thank NCSU and all the sponsors for having this program. We would also like to thank Dr. Li and Dr. Ito for their help and guidance in working on the problem. We would especially like to thank Dr. Chen for allowing us to work on this project and all the help he gave us during the workshop and other members of side-scan sonar team at Summus, Jiangying Zhou and Dohyun Chang, and Guillermo Sapiro from University of Minnesota, a consultant for the team, for fruitful discussion in formulating the problems, and to thank the Office of Naval Research for sponsoring the team. We also thank the Naval Oceanography Office (NAVO) for providing side-scan sonar images.

Bibliography

[1] J. Zhou, Y. Chen, D. Chung, G. Sapiro *Seafloor Mapping Using Side Scan Sonar Images* Technical Report. Summus Inc., Dec 2001.

[2] H. E. Reineck and I. B. Singh,*Depositional Sedimentary Environments , Springer-Verlag*, Berlin, 1975.

[3] William T. Freeman and Edward H. Adelson,*The Design and Use of Steerable Filters* , IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 13, No. 9, pp. 891-906, Sept., 1991.

[4] E. P. Simoncelli, W. T. Freeman, *The Steerable Pyramid: A flexible architecture for multi-scale derivative computation* Proc. 2nd IEEE International Conf. on Image Processing (ICIP'95), Washington, D. C., October, 1995.

[5] http://www.cis.upenn.edu/ eero/steepyr.html.

Report 6

Surface profile of granular material around an obstacle

Suhail Ahmed[1], Robert Buckingham[2], Cory Hauck[3], Christopher Kuster[4], Maša Prodanović[5], Valentin Silantyev[6]

Problem Presenter:
Tony Royal
Jenike and Johanson, Inc.

Faculty Consultant:
Pierre A. Gremaud[7]

Abstract

Closed and open flow corrective inserts are used in containers handling bulk materials in many industries to improve the flow pattern upon discharge. During the filling process, an angle of repose is formed around the fill point and an interesting free boundary problem occurs when the filling material has to flow around an insert. We will attempt to solve this problem deriving and solving a differential equation to describe the surface and investigate various algorithms similar to ray tracing techniques that might give a good approximation to the solution. The goal is to end up with the fastest algorithm that can be implemented in C code as part of a larger program used for analysis of solids flow.

6.1 Introduction and Motivation

The goal of our work is to find the surface profile of a granular material that is being poured into a hopper equipped with flow corrective inserts. Our motivation in doing this is to provide a cost effective method for measuring the volume of grain in the hopper.

The system to be modeled consists of an arbitrarily shaped hopper with arbitrarily shaped flow corrective inserts, and grain being poured in from an arbitrary point above the hopper. In modeling this system, we make several assumptions. First, we assume that the flow of the grain is slow enough that any kinetic energy

[1]Utah State University
[2]Duke University
[3]University of Maryland
[4]North Carolina State University
[5]State University of New York at Stony Brook
[6]Northeastern University
[7]North Carolina State University

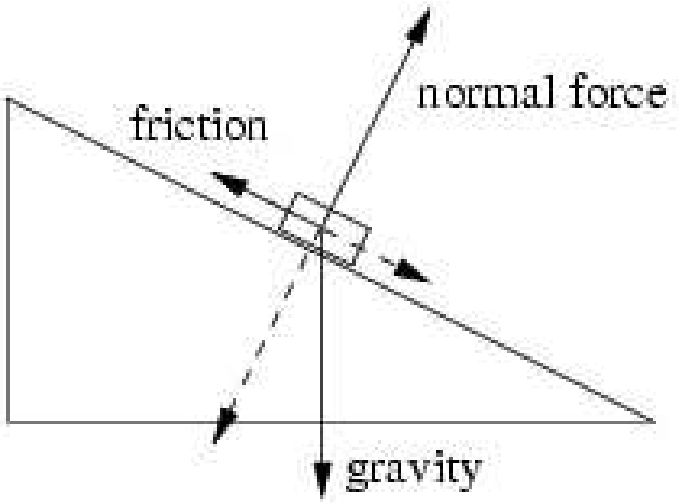


Figure 6.1: Angle of Repose: Balance of Forces

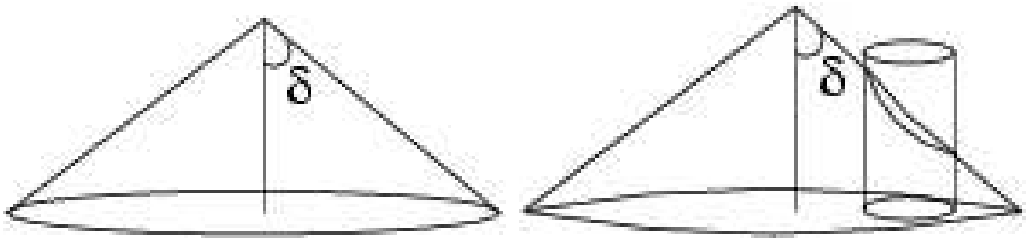


Figure 6.2: Surface Profile: No Obstruction (left), With Obstruction (right)

of the grain can be ignored. This allows us to consider the surfaces at each time to be solutions to a static force equation. This leads to surfaces where the maximum angle of incline is equal to the angle of repose, δ, for the granular material. The angle of repose is the angle at which the forces of friction and gravity cancel out for a particle on the surface of the heap (See Figure 6.1). The solution in a region containing no obstacles is a cone with angle equal to the angle of repose (See Figure 6.2). This angle is maintained even around obstacles. This leads to a shadowed area "downstream" from the obstacle where the height at a given point is lower than if the surface were a cone (See Figure 6.2).

Two methods were used to find the surface profile. The first is analogous to optical ray tracing techniques used in computer graphics. The height, $z = h(x, y)$, of the surface at any point is found by multiplying the slope given by the angle of repose by the length of the shortest path between the source and that point. For simplicity, we only considered prismatic inserts. This reduced the problem to finding the shortest path between each point and the source in the $x - y$ plane projection. This method is elaborated on in Section 6.2.

The second method used is the Fast Marching Method which is based on entropy-satisfying upwind schemes and fast sorting techniques. This method is commonly used to solve a variety of static Hamilton-Jacobi equations. The condition that the maximum angle of incline is equal to the angle of repose makes solving this problem equivalent to solving the Eikonal equation $\|\nabla T\| = F$, where $F = \tan(\delta)$ is constant. The Eikonal equation is in the Hamilton-Jacobi class of equations, and therefore the Fast Marching Method is well suited to this problem[14]. Section 6.3 describes this method in more detail.

6.2 Ray Tracing Algorithm

Ray tracing in computer visualization is a method of producing realistic images, in which the paths of individual rays of light are followed from the viewer to their points of origin. In the light ray tracing process, the physics

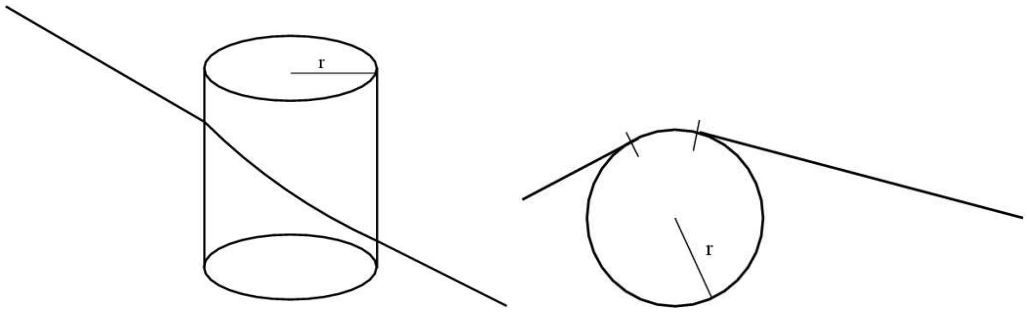


Figure 6.3: Shortest Path Around Obstacle

and mathematics of light is used. Though light rays are not being modeled in this problem, this method is called a ray tracing method since a similar approach is used.

6.2.1 Why does shortest distance in x-y plane work?

It is observed that during the filling process, non-cohesive granular material discharged from a single source will always form a surface that makes a constant angle with respect to the vertical. This angle is called the angle of repose, and the surface of the granular material is well approximated by a cone. The whole filling process can be approximated as a sequence of cones that build atop of each other as more volume flows in from the source.

What happens in the case when there is an obstacle? The answer lies partly in the idea of visibility. Two points are called visible if the line segment between them does not intersect an obstruction. The assumption of constant angle of repose can be written as

$$|\nabla h| = \tan\delta$$

wherever $\nabla \equiv (\partial_x, \partial_y)$ is defined. It is assumed that any path on the surface that follows the height gradient (in the positive direction) will not intersect the interior of an obstruction. This requires that at any point on the surface of an obstacle

$$\left(\nabla h, |\nabla h|^2\right) \cdot \nu \geq 0$$

where ν is any outward normal to the surface of the obstacle. If the obstruction is smooth, then ν will be unique, and the above requirement will be strict for almost every point on the obstacle surface that is visible from the source. This is because flow lines do not have to avoid the obstacle. However, if a point is not visible to the source, the flow lines will have to move around the obstacle tangent to its surface. In this case, the requirement above holds as an equality.

Both assumptions are physically very natural and with them, one can show that the height at any point on the surface is determined using shortest paths (See Figure 6.3). To this end, let $\mathcal{P}$ be the set of all possible paths between a point $P = (a, b, h(a, b))$ and the source $S = (0, 0, h(0, 0))$ along the surface of granular material. *Possible* paths are those with end-points at P and S that do not intersect themselves and that do not intersect the interior of any obstacle. Denote a typical path in $\mathcal{P}$ by Γ, and the projection of any path Γ onto the x-y plane by $\bar{\Gamma}$. Finally let $\mathcal{Q} = \{\bar{\Gamma} : \Gamma \in \mathcal{P}\}$, that is, $\mathcal{Q}$ is the set of all x-y projections of path in $\mathcal{P}$.

By the assumptions above, $|\nabla h| = \tan(\delta)$ wherever h is differentiable. Moreover, if we trace backward from P along the surface in the direction of the height gradient, we will never intersect an obstacle. Thus the height along that path will increase at a constant rate until we reach the height of the source. Since the source is the only point on the surface at that height, the other end of the path must in fact *be* the source. This means that there exists at least one path Υ such that $\Upsilon \in \mathcal{Q}$ and

$$\int_\Upsilon \nabla h \cdot d\vec{\ell} = \int_\Upsilon |\nabla h|\, ds = \int_\Upsilon (\tan(\delta))ds$$

where $d\vec{\ell}$ is the differential oriented path length, $ds = \left|d\vec{\ell}\right|$, and the integral along Υ is from P to S.

We claim now that

$$\text{length}\left(\Upsilon\right) = \min_{\bar{\Gamma}\in\mathcal{Q}} \text{length}\left(\bar{\Gamma}\right)$$

Suppose that this is not the case, that there exists a path $\Upsilon' \in \mathcal{Q}$ such that length(Υ') <length(Υ). Even so,

$$h\left(0,0\right) - h\left(a,b\right) = \int_{\Upsilon'} \nabla h \cdot d\vec{\ell} = \int_{\Upsilon} \nabla h \cdot d\vec{\ell} = \int_{\Upsilon} (\tan(\delta)) ds$$

Therefore if length(Υ') < length(Υ), then at some point (x, y) on Υ'

$$\left|\nabla h\left(x,y\right)\right| > \tan(\delta)$$

However, this contradicts the original assumption and means that the angle of the surface with vertical was smaller than the angle of repose. Thus the surface is too steep to be in static equilibrium and the granular material has to slide down to a new position.

Now the formula above can now be used to compute the height drop between S and P :

$$h\left(0,0\right) - h\left(a,b\right) = \int_{\Upsilon} (\tan(\delta)) ds = \tan(\delta) \times \text{length}\left(\Upsilon\right)$$

where Υ is a shortest path in $\mathcal{Q}$ between $(0,0)$ and (a,b). Therefore the ray tracing algorithm focuses on finding the shortest unobstructed path between a given point and the source.

Obstacles are assumed to be prismatic, i.e., cross sections parallel to the $x-y$ plane have no variation along the vertical axis. The reason for this restriction is the difficulty in determining visibility for a fully three dimensional obstruction. In this case, visibility depends on the vertical position, thus affecting the possible paths. The set of possible paths determines the minimum distance traveled which in turn, affects the vertical position. The coupling of planar and vertical motion in this way makes implementation of a shortest distance algorithm very difficult, especially when considering the practical restrictions imposed by the underlying mesh.

6.2.2 Outline of the Algorithm

Assumptions and remarks:
Obstacles are assumed to be right prisms with every cross section parallel to x-y plane identical. The base of the prism can be any polygon. Let a_0 be the source representing the fill point, and let a_1, ..., a_n be polygonal vertices of the obstacle base in either clockwise or counterclockwise order. Points are characterized by their coordinates and a tag called 'source' which is equal to one if the point can be a source (with respect to the point calculated at the moment, see Concavity test) and zero otherwise. The z coordinate of each point will ultimately contain either its height on the surface we are trying to compute or zero if it is inside the obstacle.

Helpful routines:
▷ Computing visibility between an arbitrary point in the plane and the point in the array a (in a there are source and vertices of the polygonal base of the obstacle). If the arbitrary point is outside the obstacle, the routine reduces to checking whether line of sight crosses any of the edges in polygon. If computing the visibility between the two polygonal vertices then concavity/convexity of the polygon at those points has to be taken into account.
▷ Determining whether a point is inside a polygon or not.

Input:
▷ Source position (in 3D) and positions of the obstacle polygonal base vertices (in 2D).
▷ Slope = tan(δ) where δ is the angle of repose.

Initializing:

1. Determine the visibility between all pairs a_i, a_j for $i, j = 0 \ldots n$.

2. Initialize the list *Sources* with original source a_0.

3. Heights of all polygonal vertices are set to 0 and the height of the source is given.

Main Loop:
While (list *Sources* not empty) do the following:

1. Let a_i be the first element in the list *Sources*
2. For all $a_j,\ j = 0..n$ in array such that a_j is visible to a_i calculate
   $height = height(a_j) + distance(a_i, a_j) * Slope$
   If $(height > height(a_i))$ then
   set $height(a_i) = height$
   if node a_j satisfies *Concavity test* add it to the end of the list *Sources*
3. Go to 1.

Assume we have vertex a_i and vertex a_j that is visible from a_i. Then *Concavity test* mentioned in the algorithm simply determines whether vertex a_j (whose height is bigger than the height of a_i) can be a source for a_i, that is can the path of sight from a_0 to a_i go through a_j. That decision at this point depends on the geometry(or rather concavity) of the polygon. For instance, a_j in the Figure 6.4 cannot be a source for a_i, but a_i could be a source for a_j.

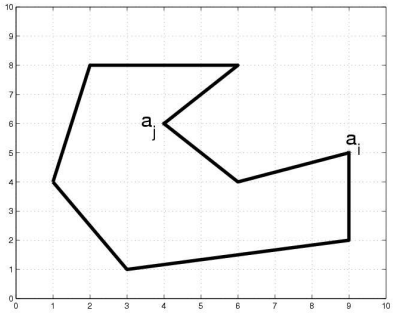


Figure 6.4: Concavity test

6.3 Fast Marching Method

6.3.1 Theory

Fast Marching Methods, invented by J.N. Tsitsikls [17, 18] and thoroughly investigated by J. A. Sethian [11, 14, 15], are numerical schemes used to compute solutions for a large class of Hamilton-Jacobi equations. Of particular interest is the nonlinear Eikonal equation defined on a set $\Omega \subset \mathbb{R}^2$ with known values on $\Gamma \subset \partial\Omega$

$$|\nabla T(x, y)| = F(x, y), \quad (x, y) \in \Omega$$

$$T(x,y) = f(x,y), \quad (x,y) \in \Gamma$$

If the surface of a cone with angle of repose δ is described by a height function

$$z = h(x,y), \quad (x,y) \in \Omega$$

$$h(0,0) = h_0$$

then h satisfies the Eikonal equation with constant $F(x,y) = \tan(\delta)$ and boundary data h_0 on $\Gamma = (0,0)$. If the origin of the $x-y$ plane corresponds to the fill point where the apex of the cone is formed), then level curves of h are the set of all points equidistant from the origin in the x-y plane so that a contour plot of h consists of concentric circles centered around the origin. In a more general setting, F may not be constant Γ may be more than just a single point, and a contour plot is not be so trivial.

Another point of view is to consider the origin as the source for a disturbance that propagates radially. Thus $T(x,y)$ measures the arrival time for (x,y), i.e., the time it has taken for the disturbance to propagate from the source to that point in the plane. When $F(x,y) = \tan(\delta)$ the arrival time corresponds directly to height drop from the top of the cone. Qualitatively, the picture for more general F and boundaries is easier to understand from this point of view. Level curves of T represent points whose arrival time from the origin is the same, and the speed at which a disturbance propagates can depend on the medium through which it passes so that F has a spatial dependence. A larger value of F at a given point in the plane is corresponds to a *slower* propagation speed. Letting $F = \infty$ corresponds then to an obstacle through which the front cannot pass. This is key to our application.

The qualitative picture for more general boundaries Γ is more difficult and is representative of the typical problem that any scheme must overcome – that solutions in general will not be differentiable even if the boundary data is smooth. Although the Eikonal equation is a boundary value problem, the lack of differentiability is analogous to the development of shocks in hyperbolic evolution equations. One can imagine, for example, the case of a propagating fire line [16] in which two separate fires merge to form a non-convex curve. At this point it is not clear mathematically how the fire-line will evolve since $t$he question of arrival time requires knowing from where the disturbance is coming. Mathematically, there are many solutions. (Consider, for example, the Eikonal equation with constant right hand side in one dimension. Any saw-tooth wave with appropriate slope and boundary value will be a solution). Thus conditions must be set that extract a unique solution. One way to do this is to introduce an entropy condition of textit first arrival time. In terms of the fire line, this means that once something burns, it can't be burnt again; in optics, this idea is related to the well known Huygens principle of light; in mathematics, the first arrival time solution corresponds to the viscosity solution of the Eikonal equation, obtained in the limit as ϵ goes to zero of the sequence of unique solutions to

$$|\nabla T(x)| = F(x) + \epsilon \nabla^2 T(x)$$

The viscosity solution is known to be the unique physically relevant, entropy satisfying solution to the Eikonal equation [4]

Having set a uniqueness criterion, the question then is how to implement a consistent numerical recipe. The idea is to use an upwind scheme that mimics the propagation of a front that at each iteration represents a level curve of T. To this end, points are evaluated in a thin band which forms the boundary between grid points who values are known at a certain iteration, and those which are not. Arrival times (and hence height values) are then computed in such a way that respects the first arrival condition. This means that the arrival time of a point P is approximated using, in both the x and y direction, the arrival time of its nearest neighbor deemed most believable, that is, the neighbor having the shortest determined arrival time. In the context of our application, this means the greatest height.

6.3.2 Algorithm

To begin, assume the obstacle is a right-angle prism. In this case, the problem reduces to solving the Eikonal equation in two dimensions. This method can be extended to a non-prismatic obstacle [see below].

Assumptions and Remarks:
We are working on two-dimensional square grid. Points are characterized by their coordinates, T-value, and

tag. At each step of the algorithm, each point on the grid is tagged as a BOUNDARY, OBSTACLE, ALIVE, CLOSE or FAR point. Points on the boundary and obstacle will always be BOUNDARY or OBSTACLE points, respectively. At each stage of the algorithm, FAR points are those which the algorithm has not incorporated yet. A point is ALIVE if its value of T is permanently set to a finite number. When computing the T-value at a point only values from its ALIVE neighbors can be used. A list of CLOSE points is maintained separately.

Input:

- ▷ BOUNDARY and OBSTACLE points.
- ▷ source position (or initial condition points) and the corresponding value(s) of T
- ▷ $\tan(\delta)$ where δ is the angle of repose .

Initializing:

1. BOUNDARY and OBSTACLE points are tagged as such. Their values are irrelevant and set to infinity.
2. The initial condition points are tagged as ALIVE and their values are set to the given values. Initial CLOSE points are the points that are neighbors of ALIVE points where square grid neighbors of $u_{i,j}$ are $u_{i-1,j}, u_{i+1,j}, u_{i,j-1}$ and $u_{i,j+1}$. CLOSE points values are computed according to the upwinding formula given in the next section. All other points have the FAR tag and their values are set to infinity.

Main Loop:
While (the list of FAR points non-empty) do the following:

1. Let TEST be the point with the smallest value among the current CLOSE points.
2. (Re)tag as CLOSE all of the FAR neighbors of TEST and add them to CLOSE list.
3. Recompute the values of all CLOSE neighbors of TEST.
4. (Re)tag the TEST point as ALIVE.
5. Go to 1.

This algorithm computes level sets of T. Since the angle of repose for a given material remains constant, the height of the sand pile is proportional to T.

6.3.3 Two-dimensional implementation

The key to this algorithm is step three. We use a first order finite difference approximation of the gradient:

$$|\nabla T| = [\max(D_{ij}^{-x}T, -D_{ij}^{+x}, 0)^2 + \max(D_{ij}^{-y}T, -D_{ij}^{+y}, 0)^2]^{1/2}$$

where

$$D_{ij}^{-x}T = T_{i,j} - T_{i-1,j}$$

and

$$D_{ij}^{+x}T = T_{i+1,j} - T_{i,j}$$

The Eikonal equation becomes an equation for T_{ij}. Since T_{ij} is a variable, it is not possible to decide which quantity to use for the maximum *a priori*. Instead, we solve eight different equations corresponding to the different combinations (disregarding the zero-zero combination) and use the equation which gives the smallest value of T_{ij}. Since the T value of FAR and BOUNDARY points is infinite, they will not contribute to the calculation. Once the T values for all points are computed, the height of the sand pile at (i, j) is given by $h - T_{ij}$ where h is the height at the source.

The two-dimensional algorithm, as is, will not work with non-prismatic obstacles, since at different T-values (corresponding to different heights in the sandpile) the cross-section of the obstacle with the x-y plane will be

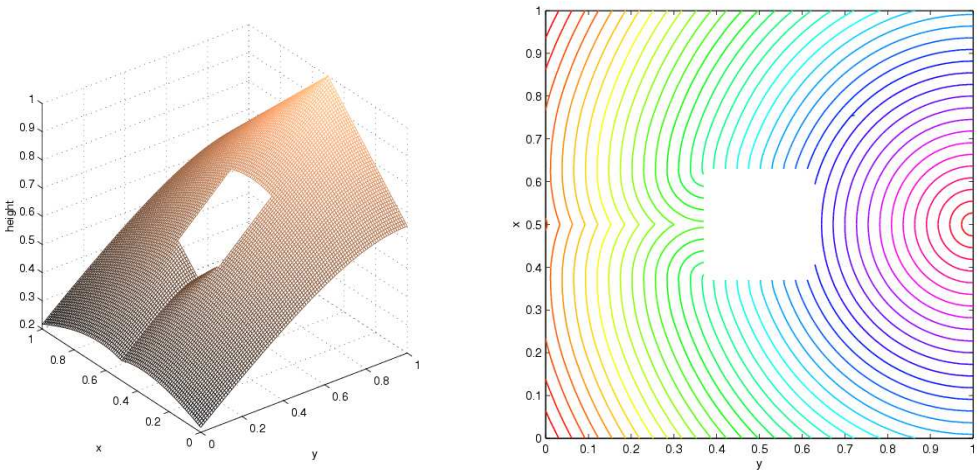


Figure 6.5: Ray Tracing/Fast Marching: Square Prism

different. For example, the cross-section of a circular cone at different T-values will be circles of increasing or decreasing radii. To overcome this difficulty the obstacle is not fixed at the beginning. Instead, the obstacle shape is parameterized by the T-value. When doing the gradient calculation, the algorithm checks to see if a neighbor of T_{ij} is in the obstacle (which depends on the height of T_{ij}). If so, that point is not used in the calculation. Computation continues until no CLOSE points remain. Points inside the obstacle will remain FAR for each step and will not be computed.

6.4 Main Results

6.4.1 Results of Computation

Ray Tracing Algorithm

Below are plots of the calculations for a square prism (Figure 6.5), a non-convex prism (Figure 6.6), an asymmetric prism (Figure 6.7), and a circular prism (Figure 6.8). The results for the first two shapes are exact, but the results for the circular prism are not since the boundary was approximated by a 20 vertex polygon.

Fast Marching Method

Below are plots of a square prism (Figure 6.5), a circular prism (Figure 6.8), and a non-convex prism (Figure 6.6). The third shape shows the algorithm can handle concavities. The fourth demonstrates non-symmetric obstacles (Figure 6.7). These four are indistinguishable from the results of the ray tracing method. However, this algorithm is also capable of handling disconnected obstacles (Figure 6.9) as well as non-prismatic obstacles (Figure 6.10).

6.4.2 Limitation and Difficulties

Ray Tracing Algorithm

Once the shortest paths are found, the calculation is straightforward and as accurate as the floating point representation on computers. The only accuracy problem comes from how well the obstacle base can be approximated by a polygon and using too big number of polygonal vertices can eventually slow down the calculation. It is more or less straightforward to implement the method for more than one obstacle as long as the problem can be reduced to two dimensions.

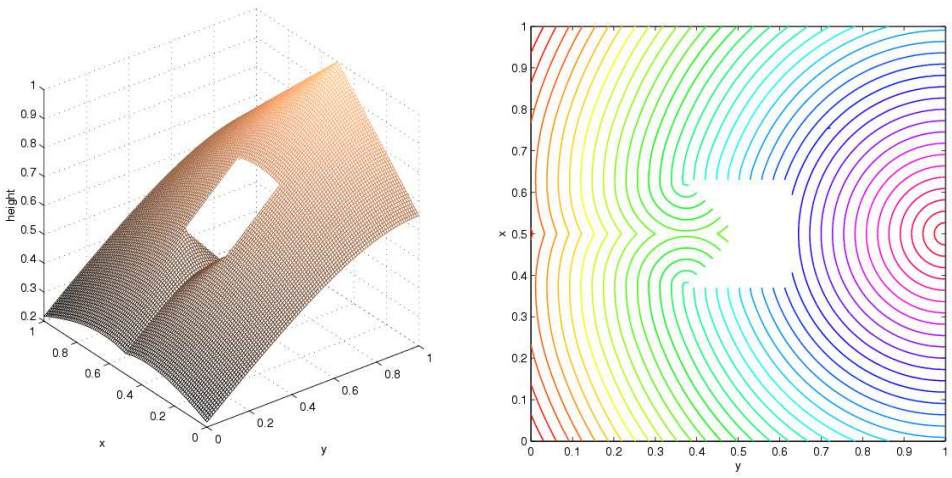


Figure 6.6: Ray Tracing/Fast Marching: Prism with Downstream Concavity

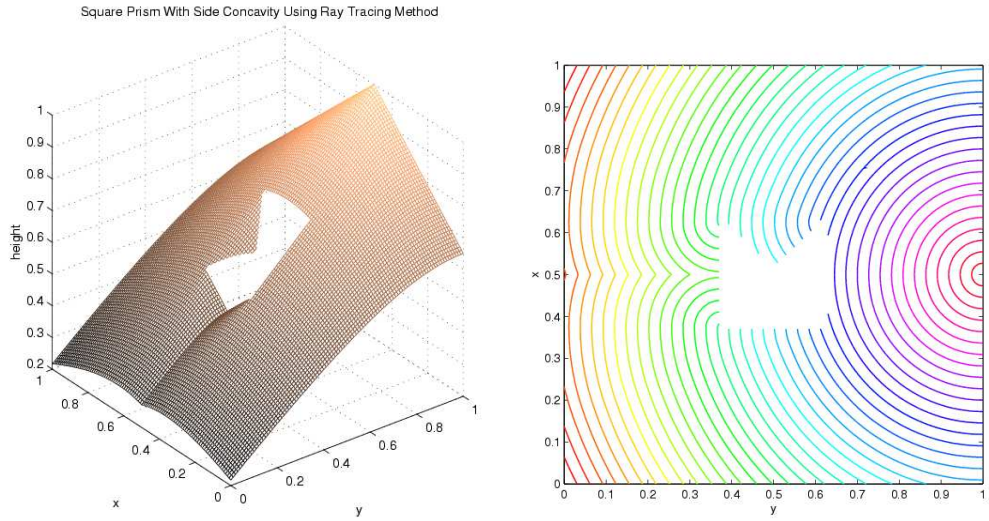


Figure 6.7: Ray Tracing/Fast Marching: Asymmetric Prism with Concavity

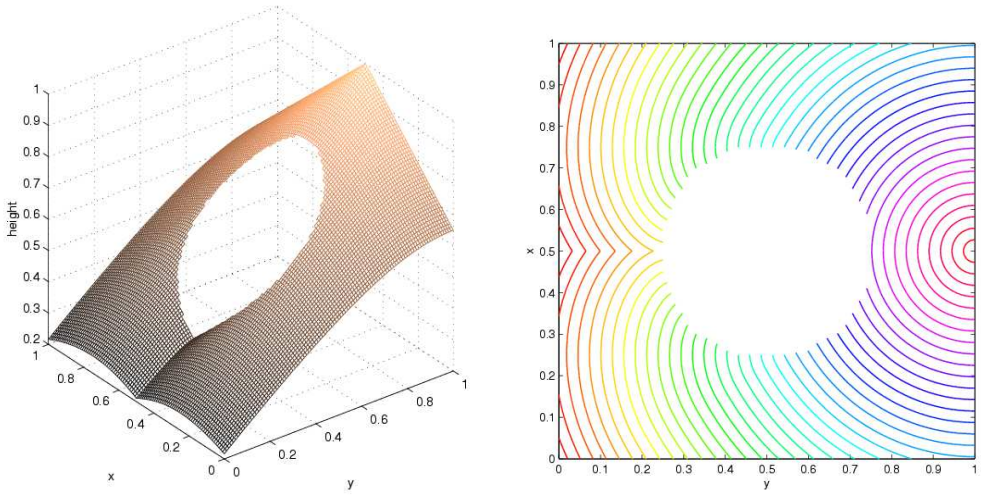


Figure 6.8: Ray Tracing/Fast Marching: Circular Prism

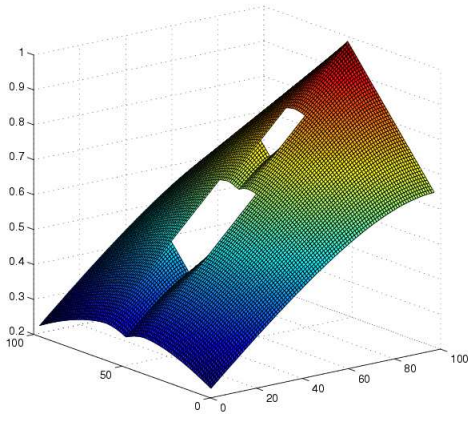


Figure 6.9: Fast Marching: Disconnected Obstacle

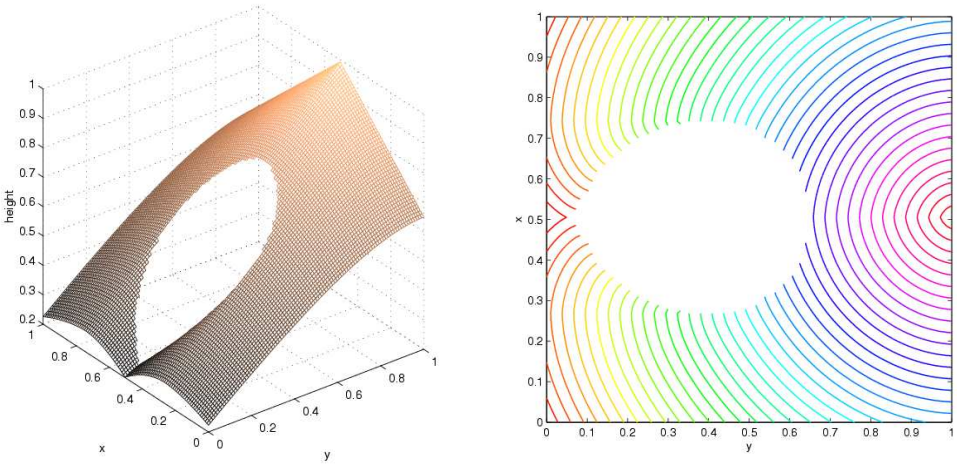


Figure 6.10: Fast Marching: Non-Prismatic Obstacle (Cone)

However, it is computationally very expensive to implement the method for obstacles with cross sections that change with z-coordinate since such implementation requires efficient computation of visibility between all pairs of vertices.

Fast Marching Method

The two-dimensional algorithm works for any prismatic obstacle, including those with asymmetries and concavities. There could be multiple, disconnected obstacles (i.e. two cylinders). It is necessary to modify this algorithm in a straight-forward way to deal with truly three-dimensional obstacles. The method works for simple objects such as cones and pyramids. However, there are several types of obstacle for which the algorithm will fail. The method must be modified before it could predict sandpile shapes around obstacles with holes. There will also be a problem with cliffs (where the sand will build up on top and eventually fall off). This has the effect of creating a second source. Currently the algorithm assumes a unique source.

One disadvantage of the fast marching method as compared to the ray tracing method is that fast marching is an approximate scheme. However, it will converge to the correct answer as the mesh size is decreased. The scheme presented here is first order, but it is possible to use a higher order method for the gradient calculation.

6.4.3 Conclusions

Finding the shape of a granular heap in a hopper with corrective flow inserts is of interest in industry. For example, this would allow computation of the volume of granular material in the heap given only the geometry of the hopper and the height of the pile underneath the source. Two distinct methods, ray tracing and fast marching, have been shown to give the profile of heaps around a large class of obstacles. The only assumption about the heap is that the angle with the horizontal, or angle of repose, is constant everywhere.

The ray tracing method assumes the obstacle is prismatic, that is, depends only on x and y. First the boundary of the obstacle is discretized. It is important to keep track of which boundary points are visible from each other and from the source. To find the height of the pile at any point the length of the shortest path from to the source in the $x-y$ plane is computed. Since the angle of repose δ is constant along this path, the change in height from the source is found by multiplying the length of this path by $\tan(\delta)$. The accuracy of this method depends on how finely the obstacle boundary is discretized.

The fast marching method works for more general obstacles than ray tracing such as cones, which are of primary interest in the industry. Level sets are computed in an upwind scheme from the source using the Eikonal equation $||\nabla T|| = const.$ The height at the source is given initially. Then the nearest points are updated in an expanding fashion until every grid point has a height value. The accuracy of this method depends on the order of the scheme used to compute the gradient.

Acknowledgments

We would like to thank Tony Royal and Jenike and Johanson, Inc. for supplying us with this problem.

Bibliography

[1] P. Bose, A. Lubiw, and J.I. Munro, *Efficient Visibility Queries in Simple Polygons,* Computational Geometry, accepted December 2001.

[2] D.L. Chopp, *Computing Minimal Surfaces Via Level Set Curvature Flow,* J. Comput. Phys., 106 (1993), pp. 77-91.

[3] M.G. Crandall and P.L. Lions, *Viscosity Solutions of Hamilton-Jacobi Equations,* Trans. Amer. Math. Soc, 277 (1983), pp. 1-43.

[4] M.G. Crandall, L.C. Evans, and P.-L. Lions, *Some properties of viscosity solutions of Hamilton-Jacobi equations,* Trans. Amer. Math. Soc., 282 (1984), pp. 487-502.

[5] P.W. Cleary and M.L. Sawley, *DEM Modeling of Industrial Granular Flows: 3D Case Studies and the Effect of Particle Shape on Hopper Discharge,* Appl. Math. Model, Vol. 26, No. 2 (2002), pp. 89-111.

[6] E.W. Dijkstra, *A Note on Two Problems in Connection with Graphs,* Numer. Math., 1 (1959), pp. 269-271.

[7] Y. Grasselli and H.J. Herrmann, *Shapes of Heaps and in Silos,* European Physical Journal B, 10 (1999): pp. 673-679.

[8] H.J. Herrmann, *Simulation of Granular Media,* Physica A, 119 (1992): pp 263-276.

[9] H.J. Herrmann, *Shapes of Granular Surfaces,* Physica A, 270 (1999): pp. 82-88.

[10] D.V. Khakhar, A.V. Orpe, P. Andresen, and J.M. Ottino, *Surface Flow of Granular Materials: Model and Experiments in Heap Formation,* Journal of Fluid Mechanics, 441, (2001): pp. , .

[11] R. Kimmel and J.A. Sethian, *Computing Geodesic Paths on Manifolds,* Proc. National Academy of Sciences, July 1997.

[12] R.W. Lyczkowski, I.K. Gamwo, and D. Gidaspow, *Analysis of Flow in a Non-Aerated Hopper Containing a Square Obstacle,* Powder Technology, 112 (SI October 2000): pp. 57-62.

[13] R.M. Nedderman, S.T. Davies, and D.J. Horton, *The Flow of Granular Materials Round Obstacles,* Powder Technology, Vol. 25, No. 2 (1980): pp. 215-223.

[14] J.A. Sethian *Fast Marching Level Set Methods for Three-Dimensional Photolithography Development*, SPIE, v2726, pp261-272, 1996.

[15] J.A. Sethian *Fast Marching Methods,* SIAM Review, Vol. 41, No. 2, 299-235, 1999.

[16] J.A. Sethian, *An Analysis of Flame Propagation,* Ph.D. Thesis, Dept. of Mathematics, Univ. of Berkeley, 1982.

[17] J.N. Tsitsikls, *Efficient Algorithms for Globally Optimal Trajectories*, Proceedings of IEEE 3rd Conference on Decision and Control, Lake Buena Vista, Fl, Dec 1994.

[18] J.N. Tsitsikls, *Efficient Algorithms for Globally Optimal Trajectories*, IEEE Transactions on Automatic Control, v40, pp1528-1538, 1995.

Report 7

Predictive toxicology: benchmarking molecular descriptors and statistical methods

Predictive Toxicology: Benchmarking Molecular Descriptors and Statistical Methods

Jun Feng[1], Laura Lurati[2], Haojun Ouyang[3], Tracy Robinson[4], Yuanyuan Wang[5], Shenglan Yuan[6], S. Stanley Young[7]

Abstract: The development of drugs depends upon finding compounds that have beneficial effects with a minimum of toxic effects. The measurement of toxic effects is typically time consuming and expensive so there is a need to be able to predict toxic effects from compound structure. In this paper, combinations of different chemical descriptors and popular statistical methods were applied to the problem of predictive toxicology. Predicting toxic effects is expected to be challenging, as there are usually multiple toxic mechanisms involved. Four datasets were collected and cleaned, and four different sets of chemical descriptors were calculated for the compounds in each of the four data sets. Three statistical methods, recursive partitioning (RP), neural networks (NN), and Partial Least Square (PLS), were use to attempt to link chemical descriptors to the response. Good predictions were achieved in the two smaller datasets; we found for large datasets the results were less effective, indicating new chemical descriptors or statistical methods are needed. All the methods and descriptors worked to a degree but our work hints that certain descriptors work better with specific statistical methods so there is a need for continued methods understanding and development.

Introduction

The development of drugs depends upon finding compounds that have beneficial effects with minimal toxic effects. Currently we do not have the knowledge and computational power to directly predict toxic effects, so we have to rely on structure-activity relationship, SAR, methods. Unlike initial screening and lead optimization, in which the binding affinities to a specific receptor are typically modeled, toxicity often involves many different receptors and mechanism, but with the same toxicology response, death, mutagenicity, etc.

Currently there are two predictive toxicology strategies: knowledge-based and statistically-based. Knowledge-based methods rely on rules from human experts or previous knowledge on structures-toxicity relationships, while statistics-based methods rely on using training data sets, chemical descriptors, and statistical methods to correlate with the observed toxicity data to structural features, and generate a mathematical predictive model. The predictive model is used to evaluate untested compounds. There are a number of commercial available software packages for making predictions of toxic effects: DEREK, ONCOLOGIC, HAZARDEXPERT, TOPKAT, etc. DEREK,

[1] University of North Carolina at Chapel Hill
[2] Brown University
[3] University of Toledo
[4] North Carolina State University
[5] University of Waterloo
[6] Auburn University
[7] CG Stat.

ONCOLOGIC and HAZARDEXPERT are knowledge-based and TOPKAT is statistically based.

Generally, multiple linear regression is the dominant statistical method used for predictive toxicology. Since most toxicology prediction involves heterogeneous compound classes and multiple toxic mechanisms, we expect more complex statistical methods like recursive partitioning, RP, neural networks, NN, and partial least squares, PLS, should perform better than linear regression. We are also curious that certain types of chemical descriptors might work better with certain types of statistical analysis so we study the performance of different chemical descriptors with different statistical methods. We want our findings to have some generality so our study uses four different data sets.

Our goal is to investigate the performance of three statistical methods, Partial Least Squares, Neural Networks, and Recursive Partitioning (PLS, NN, and RP) and four types of chemical descriptors, Constitutional, Topological Index, BCUT, and Fragment/Property descriptors, (Cons, TI, BCUT, Frag). The statistical methods and molecular descriptors are described in more detail later. We are interested if certain types of descriptors work better with specific statistical analysis methods. For example, do TIs work better with PLS and fragments work better with NN? The three statistical methods we used are advertised to work well in complex situations. Partial Least Squares (PLS) is capable of dealing with the situation where there are more descriptors than observations. Neural Networks (NN) is capable of dealing with nonlinear relationships. Recursive Partitioning (RP) is capable of dealing with multiple mechanisms. We use four publicly available datasets (three toxicology data sets and one potency data set). We calculated four sets of chemical descriptors using free-ware DRAGON. These data sets, SD files and descriptors, are posted at **www.niss.org**. Altogether there are 48 situations, four datasets, four sets of chemical descriptors and three statistical methods.

This research followed the following plan. First, four public data sets were obtained. The point here is to determine the effectiveness of the chemical descriptors and statistical methods over a variety of data sets. Second, four classes of descriptors were computed for each data set. There is considerable interest in molecular descriptors and, again, we wanted to access the different statistical methods using different descriptors. Third, we used three popular statistical analysis methods, PLS, NN and RP. Each data set was divided at random into two data sets, a training data set for training the model and testing data set for testing the predictions of the model. A prediction equation was made using a statistical method, a type of molecular descriptor, and the training data. The quality of the prediction was evaluated using the testing set. As evaluation of all statistical methods on all data sets and descriptor sets would require more labor than available, we used a statistical sampling plan to select 15 of the 48 combinations of data set/descriptor type/statistical methods. Three of the 15 conditions were replicated by re-splitting the data at random to give a new training and testing data set. Examination of the replicate evaluations can give a sense of the variability of the results.

This paper is organized as follows. First, we describe the data sets and data processing. Second, we describe four sets of chemical descriptors. Third, we describe the three

statistical methods. Frag, we then use statistical design of experiments to select an informative subset from the 48 possible combinations of statistical method, descriptor type, and data set. We give the results and conclude with a discussion.

Methods (Data, Descriptors, and Statistical Methods)

Data and Data Processing

The data sets used are listed in Table 1. There are several tasks involved in data processing. The logistical order of these tasks is

1. Collect data sets
2. Pre-format data for compatibility with Dragon
3. Compute descriptors
4. Post-format data for compatibility with ChemTree and JMP .

The descriptors were computed using Dragon, developed by Todeschini, Consonni, and Pavan of Milano Chemometrics. The descriptors used were separated into four sets listed in Table 2. These descriptors are described in Todeschini and Consonni (2000) and can be viewed at http://www.dist.umimib.com. Dragon is limited to processing 1,500 compounds at a time, so the files were divided, processed and reconstituted. Incompatible electronic compound representations were also resolved or the compound was eliminated. Getting and cleaning chemistry data sets for descriptor calculation and statistical analysis is difficult and time consuming; our team spent five days cleaning the data sets, computing the molecular descriptors, and staging the data for statistical analysis.

Chemical descriptors

All descriptors were computed from SD files using Dragon.

Constitutional Descriptors [Todeschini and Consonni, 2000, pages 90-91]**:** There are 47 descriptors. Examples include molecular weight, atomic weight, atom counts, etc. Descriptors in this class are not determined by the connectivity or conformation of the molecule.

Topological Information Indices [Todeschini and Consonni, 2000 pages 447-456]**:** Topological Index, TI, descriptors are widely used in QSAR analysis because they are easy to calculate, do not depend on conformation, and are sensitive to small changes in molecular structure. Dragon can calculate 262 kinds of TIs. Although sometimes very good regression models can be obtained using TIs, they remain controversial, as it is essentially impossible to associate chemical structural meaning to them.

BCUT [Burden, 1989; Burden, 1997; Todeschini and Consonni, 2000 pages 132-134]**:** Similar to topological index, BCUT descriptors are also determined by the connectivity, i.e. topological relations between different atom within the molecule. BCUT descriptors are calculated from the adjacent matrix, also called Burden matrix. In this matrix, atomic properties are placed on the major diagonal and a measure of connection is placed in the

off-diagonal cells. Through diagonalisation of the adjacent matrix, eight highest eigenvalues and eight lowest eigenvalues are used. The four atomic properties are atomic mass, van der Waals volume, atomic electronegativitie, and atomic polarizability, so there are 4X16=64 descriptors. Like TIs, BCUTs are controversial, as their utility is not firmly established for QSAR and it is essentially impossible to associate chemical structural meaning to them.

Fragment/Property Descriptors [Todeschini and Consonni, 2000]**:** We also tried to use some descriptors that reflect the physico-chemistry properties, like logP, aromatic index, etc. Fragment descriptors, which indicate what kinds of fragments are in the molecule and how many of them, are also included in this class.

Table 2 lists the number of descriptors in each class and also the number of eigen values greater than 1 for each class as an indication of the number of real variables in each class. We were surprised at the relatively small number of greater than one eigen values for each class and particularly surprised that there were only four eigen values greater than one for BCUTs.

Statistical Methods

The statistical programs used for this analysis were JMP and ChemTree . Some of the data processing was done in SAS. ChemTree has internally computed fragement-based descriptors; these were not used for this study.

Partial Least Square

Partial least squares (PLS) is a statistical method for analysis of systems of independent and response variables. Weighted linear combinations of the predictor variables are used to predict the response variable(s). PLS is a predictive technique that can handle more independent than response variables and also can relate the set of independent variables to a set of multiple dependent (response) variables. PLS regression has been used in various disciplines such as chemistry, economics, medicine, psychology, and pharmaceutical science where predictive linear modeling, especially with a large number of predictors, is necessary. PLS regression is probably the least restrictive of the various multivariate extensions of multiple linear regression. This flexibility allows it to be used in situations where traditional multivariate methods fail, such as when there are fewer observations than predictor variables. PLS regression can be used as an exploratory analysis tool to help select suitable predictor variables and to identify outliers before classical linear regression is used.

PLS is an extension of multiple linear regression. In multiple linear regression a model specifies the (linear) relationship between a dependent (response) variable *Y*, and a set of independent (predictor) variables, the X s, so that

$$Y = b_0 + b_1 X_1 + b_2 X_2 ... + b_p X_p$$

Where b_0 is the regression coefficient for the intercept and the b_i values are the regression coefficients (for variables 1 through p) computed from the observed data.

In our case we use PLS to build a linear model on each training set, Y=XB+E, where Y is an m cases by r variable(s) response matrix, where m is the number of compounds in the training set. In our case r=1. For the yeast data we initially chose 2 principal components i.e. r=2. To simplify this paper, we restricted our reported analysis to only one response. X is an m cases by p variable predictor (design) matrix, B is a p by r regression coefficient matrix, and E is a error term for the model which has the same dimensions as Y. Usually, the variables in X and Y are centered by subtracting their means and scaled by dividing by their standard deviation.

PLS simplifies the descriptors by use of factor scores, linear combinations of the original predictor variables. There is no correlation between the factor score variables used in the predictive regression model. For example, suppose we have a data set with response variables $\boldsymbol{Y}$ (in matrix form) and a large number of predictor variables $\boldsymbol{X}$ (in matrix form), some of which are highly correlated. A regression, using factor extraction for this type of data, computes the factor score matrix $\boldsymbol{T}=\boldsymbol{XW}$ for an appropriate weight matrix $\boldsymbol{W}$, and then considers the linear regression model $\boldsymbol{Y}=\boldsymbol{TQ}+\boldsymbol{E}$, where $\boldsymbol{Q}$ is a matrix of regression coefficients (loadings) for $\boldsymbol{T}$, and $\boldsymbol{E}$ is an error (noise) term. Once the loadings $\boldsymbol{Q}$ are computed, the above regression model is equivalent to $\boldsymbol{Y}=\boldsymbol{XB}+\boldsymbol{E}$, where $\boldsymbol{B}=\boldsymbol{WQ}$, which can be used as a predictive regression model.

PLS produces the weight matrix $\boldsymbol{W}$ reflecting the covariance structure between the predictor and response variables. PLS produces a p by c weight matrix $\boldsymbol{W}$ for $\boldsymbol{X}$ such that $\boldsymbol{T}=\boldsymbol{XW}$, i.e., the columns of $\boldsymbol{W}$ are weight vectors for the $\boldsymbol{X}$ columns producing the corresponding n by c factor score matrix $\boldsymbol{T}$. These weights are computed so that each of them maximizes the covariance between responses and the corresponding factor scores. Ordinary least squares procedures for the regression of $\boldsymbol{Y}$ on $\boldsymbol{T}$ are then performed to produce $\boldsymbol{Q}$, the loadings for $\boldsymbol{Y}$ (or weights for $\boldsymbol{Y}$) such that $\boldsymbol{Y}=\boldsymbol{TQ}+\boldsymbol{E}$. Once $\boldsymbol{Q}$ is computed, we have $\boldsymbol{Y}=\boldsymbol{XB}+\boldsymbol{E}$, where $\boldsymbol{B}=\boldsymbol{WQ}$, and the prediction model is complete. We used the PLS platform in JMP for the analysis work in this paper.

Neural Networks

Historically, the model structure of neural networks (NN) was based on our understanding of a biological neuron system. The human brain is a highly connected set of neurons. A neuron has inputs from other neurons and outputs to other neurons. Like neurons, a NN model tries to determine the linear combination of the explanatory variables that can predict a target response as functions of input features (Hastie et al. 2001).

Turning to mathematics, a NN provides a way to approximate a general non-linear function. Study and use of NN has evolved to a large number of methods and algorithms

and there is less emphasis on the biological origins of the methodology. A feed-forward NN with one hidden layer is the simplest but most common form in use(e.g. Ripley 1996); This form is available in JMP and we use this form for our predictions.

Figure 1 gives a diagram of a simple, feed-forward NN with one hidden layer (3 hidden units). It has three input (explanatory) variables and a single output (response) variable. In between the input and output layers is a hidden layer. The hidden layer can have any number of units; here it has just three.

The output units, y_k s , are functions of input information,x_i s (Venables and Ripley 1999).

$$y_k = G_0(a_k + \sum_h w_{hk} G_h(a_h + \sum_i w_{ih} x_i))$$

Usually, the function G_h on the hidden layers is a logistic function defined as

$$G(x) = \frac{\exp(x)}{1+\exp(x)}$$

The appropriate regression or classification network is constructed with corresponding continuous or categorical response. The parameters a_k s ,a_h s ,w_{hk} s andw_{ih} s can be determined by minimizing some error function such as the least-square criterion:

$$R(a_k \text{ s}, a_h \text{ s}, w_{hk} \text{ s}, w_{ih} \text{ s}) = \sum_{i=1}^{n} (y_i - y_i)^2$$

Next, we comment on the training a NN. In practice, because NN often have a large number of weights the optimization of error function of $R(a_k \text{ s}, a_h \text{ s}, w_{hk} \text{ s}, w_{ih} \text{ s})$ is very computational intensive and the global minima will often over-fit the data. In effect the NN can memorize the training data set. JMP attempts to to get around excessive number of weights and over fitting by adding a penalty to the error function:

$$\text{K}(a_k \text{ s}, a_h \text{ s}, w_{hk} \text{ s}, w_{ih} \text{ s}) =$$

$$R(a_k \text{ s}, a_h \text{ s}, w_{hk} \text{ s}, w_{ih} \text{ s}) + \lambda \text{L}(a_k \text{ s}, a_h \text{ s}, w_{hk} \text{ s}, w_{ih} \text{ s})$$

where λ is called weight decay parameter; the function L will become larger with larger effects of weights. The optimization algorithm will tend to minimize the penalty function K instead of R and large value of λ will shrink the weights toward zero.

Since the function K is non-convex, there are many local minima. With fixed λ=0.01(default) the JMP randomly selects 20 (default) starting points, finds the best local minimum which have lease value of K as a final solution.

NN fitted in JMP have only one hidden layer with three hidden units (default). Many books argue that it is better to have too many hidden unites than too few. Hastie et al. (2001) recommend that the number of hidden units be in the range of 5 to 100, with the number increasing with the number of inputs and number of training cases.

Specifically, in our application we use all the JMP default options since we believe that is what most people would try first. It should provide fair comparison to the result from other methods where we use default options as well. Note that RP, NN, and PLS can all be tuned to the problem at hand. Each software vendor has set the defaults at presumably good settings. It is beyond the scope of this research to exhaustively tune the methods.

NN is touted as a very powerful learning method with widespread application in many fields. Unfortunately, it is essentially uninterruptible. There are conflicting claims in the literature on the superiority of one method or type of chemical descriptors, so it will be of interest to compare statistical methods using different types of chemical descriptors. It is obvious that we will not resolve this controversy as each statistical method can be tuned by experts and there no end of chemical descriptors that can be computed and endless ways to mix and match them to the statistical method. Our study, by necessity is limited, but it does point to a way to benchmark statistical methods and chemical descriptors.

Recursive Partitioning

Recursive partitioning, RP, is a data mining method for finding predictive patterns in large, complex data sets. RP progressively divides a data set into smaller and more homogeneous subsets. The two main purposes of using trees are to aid in identifying the underlying data structure and to predict values of future observations. A tree is a *classification tree* if the predicted variable is a category, e.g. active/inactive, and a *regression tree* if the endpoint is a continuous number, e.g. LD50. There are a number of RP algorithms; we used ChemTree which is specifically designed for analysis of chemistry data sets.

RP works from a response variable and a measurement vector **x** which contains information about an observation on many different variables ($x_1, x_2, \ldots, x_p$). In this case, the measurement vector consists of different compound descriptors. The sample that we use to create a tree is referred to as the learning or training sample.

The dataset and descriptor combinations used with RP are as follows: NCI with TI, NCI with BCUT, mutagenicity with Cons, mutagenicity with Frag, and toxicity with Frag. The training and test sets were a random selected division of each entire set. A generic tree from ChemTree looks is given in Figure 2.

In each interior node, the information consists of the splitting variable information, the number of observations in the node (n), the toxicity mean (u), the standard deviation (s), and the Bonferroni adjusted P-value for the split. The algorithm finds the best cut

points for each continuous descriptor variable to best divide the observations into homogeneous groups. It then selects among variables for the best variable. This is computer intensive. The adjusted p-value is used to help insure that the split reflect cause rather than chance.

The first step in constructing a tree is to determine the selection of splits. The basic idea is to split at each node in a way that will make the descendant nodes more homogeneous based on their toxicity value. In other words, as we split into more nodes, the observations in each node will be more similar to each other than they are to observations in other nodes. If the response variable is binomial then a chi-square test is performed to measure the dissimilarity of the two daughter groups and if it is continuous then the dissimilarity is measured by a function based on maximum likelihood equations. (See the ChemTree Manual 2002, p. 38). P-values are calculated for either of these tests and then adjusted by multiplying by the Bonferroni corrector factor, which is the number of independent variables that could actually be used to split the node. A node is terminal when the split P-values are no longer significant for any predictor variable. For the binomial toxicity response (1=toxic, 0=not), if the mean of the terminal node is greater than 0.5, then the node s activity is classified as toxic and not toxic otherwise. For the mutagenicity data set, about half the compounds are measured as mutagens so using 50% to classify a node makes sense; when the positive effect is very rare, e.g. compound screening, then we would declare a node positive with a far lower positive rate. For a continuous toxicity response, the toxicity of a terminal node is the mean value of that node. Once the tree is created, the test data set is sent through the tree so the predictive quality of the tree can be measured. For our continuous response data sets, an R^2 value is calculated for the observed (x) versus the predicted (y) values. For our binomial response data sets, a chi-square value is calculated.

$$R^2 = \_(x_{bar}-x_I)(y_{bar}-y_I)/ [\_(y_{bar}-y_I)\_(x_{bar}-x_I)]$$

$$\_^2 = \_[\_(\text{Actual-Expected})^2/\text{Expected}]$$

Design of Experiments

Now, we have three factors: data sets, molecular descriptors and statistical methods. The goal of this project is to find out main effects of descriptors and methods and interaction between them within the blocking effect of the data sets. The ideal situation is that we can figure out which descriptors or methods are consistently good or bad cross the data set and how they interact with each other. Thus, our result can be a guide to further study.

Here are the experimental factors with their levels:
Data set (4 levels) : Toxicity, Mutagenicity, Yeast, NCI
Descriptor (4 levels) : Cons, TI, BCUT, Frag
Methods (3 levels) : PLS, NN, RP

The data sets should be considered as block in that the statistical methods and chemical descriptors are run within each data set. Each data set has a specified number of compounds and the biological potencies are more likely to be measured consistently within a data set. To run a complete experimental design would require 4*4*3=48 runs. Since we had only limited resources for our data analysis and our investigation is considered preliminary, we decided to run 18 experiments. The detailed plan is as follow: we select 15 out of 48 runs so that we could estimate all main effects, and two-way interactions between descriptors and statistical methods. We effectively treat the data sets as blocks. We then selected 3 out of 15 selected runs to replicate. One replicate was selected at random within each of the three statistical methods. The 15 point design was constructed using JMP.

Below is the table of experiment runs (* replications). To form a replicate, we re-split the data at random to form a new training and testing set.

Run	Data Set	Descriptor Type	Stat Method
1*	Mut	Cons	RP
2	Yeast	Cons	NN
3	Tox	Frag	NN
4	Tox	BCUT	PLS
5	NCI	TI	RP
6	Mut	Cons	PLS
7	Tox	Frag	PLS
8	Mut	BCUT	NN
9*	NCI	Frag	PLS
10	NCI	BCUT	RP
11	Yeast	TI	PLS
12*	Mut	TI	NN
13	Mut	Frag	RP
14	Yeast	BCUT	PLS
15	Tox	Frag	RP

Results

Recursive Partitioning

Using the topological descriptors with the NCI data set resulted in a tree with eighteen terminal nodes; of which, only two were classified as being potent. In the original test data set, 301 compounds are potent while the tree only predicted that eleven compounds are potent. If we had tested each compound individually for potency, we would have a hit rate of 301/14568= 0.02066 but if we test only the ones that are predicted to be potent, we have a hit rate of 0.18182, an 8.8 fold increase in hit rate. In order to compare this method with the other methods, a two-way contingency table is created:

		Actual		
	Count Total %	0	1	
Predicted	0	14528 97.87	299 2.05	14577 99.92
	1	9 0.06	2 0.01	11 0.08
		14267 97.93	301 2.07	14568

NCI Data with Topological Descriptor
Contingency Table

The Chi-Square for this 2x2 table is 5.473.

Five of the nineteen terminal nodes are classified as potent for the tree created using the BCUT descriptors on the NCI data set. The hit rate for this set increased from 0.02066 to 0.52703 and the Chi-Square test value is 206.687. The next data we worked with was the mutagenicity data. The constitutional descriptors were used to create trees for two different random splits of test versus treatment. The first tree had 24 terminal nodes with thirteen of those nodes being classified as toxic. The hit rate increased from 0.48595 to 0.6825 and the Chi-Square test value is 119.101. The second tree had a hit rate increase from 0.49835 to 0.74848 and a Chi-Square test value of 134.090. Ten of its 23 terminal nodes are classified as toxic. The last binomial response tree was created using the Frag descriptor set on the mutagenicity data. With nine out of sixteen terminal nodes being toxic, the hit rate for this tree increased from 0.48595 to 0.75472 and has a Chi-Square test value of 197.519. The last data set is the toxicity data set which has a continuous response variable. The R-squared value for using the Frag descriptor set to create a tree on this set is 0.494835. An observed versus predicted graph is provided (Figure 3).

Partial Least Square (PLS)

Our datasets have two types of response: continuous response and binary response. For binary response, logistic regression may be a better tool, but since the goal of our study is to investigate the performance of PLS analysis, we have to convert the predicted continuous response to binary response by applying an arbitrary cutoff. The selection of the cutoff will solely depend on how good it works on training set.

There are six PLS analyses with different combinations of descriptors and methods: BCUT descriptors on Tox dataset, BCUT descriptors on Yeast dataset, CONS descriptors on Mut dataset, TOPO descriptors on Yeast dataset, FRAG descriptors on Tox dataset, and FRAG descriptors on NCI dataset.

PLS method can handle the situation when there are more descriptors than number of observations, but for most of our test cases, there are usually many more observations than descriptors. The best results were obtained with TOX dataset, which contains 270

compounds, but as the size of datasets is increasing, the result becomes less satisfactory, especially for the NCI and Yeast dataset. We suspect that those datasets are too structurally diverse and the compounds may have operated by different mechanisms. The results, -log(p-value) of association of predicted versus observed for test data sets, are included in Table 4.

Neural Network

As given in the experimental plan, a NN was applied three data sets once: Yeast data with Constitutional descriptors (YC), Toxicity data with Frag descriptors (TF), Mutagenicity data with BCUT descriptors (MB) and one data sets twice: Mutagenicity data with Topological descriptors (MT).

All of these four data sets have many explanatory variables, which is a computational disaster to training NN models. For example, even with the YC data having only 47 the descriptors, NN with only 3 hidden units have to estimate about 150 weights. It is almost impossible to fit NN in JMP . Therefore, for each data set, we calculate the first 10 principal components and used these as the inputs, which made the computation feasible. In this way we only need to estimate about 40 parameters.

For YC, TF, MB data, we randomly split data into training and test set. For MT, two random divisions were applied. So we have five data, each having the training and test sets. We used NN, built on the training set, to predict the test set and calculate R-square (for regression) or Chi-square (for classification) for linear regression between the prediction and the observation of the response on the test set.

Statistical Analysis

For each of the 18 experimental conditions, data set, chemical descriptor, and statistical method, we computed the association between the predicted and observed values for the test data set. The prediction model was constructed using the training data set. The p-value for this association was transformed using the negative log base 10. The analysis of variance for this data is given in Table 5. We see that there is an indication of significant effects when the three replicates are used to test the remaining effects. Turning to the Effects Tests we see that there are significant differences among the data sets. It is expected that data sets will differ in the ability to be modeled. One should conclude that multiple data sets should be used for benchmarking studies. There is an indication of an interaction between the chemical descriptors and the statistical methods, $p<0.0492$. The presence of an interaction is not unexpected. The interaction effect is just statistically significant, so more benchmarking should be done. If the interaction is true, the implication is that particular descriptors will work more effectively with particular statistical methods. Interactions can be examined in a number of ways. Table 6 gives the least squares means for the statistical method by molecular descriptor type. It appears that BCUT descriptors do not work well with NN and appear to work quite well for RP. It is often useful to plot the ranked values against the integers, Figure 4. Three values in the plot appear to be larger than expected, BCUT and RP (56.95), Frag and PLS (43.15) and

Frag and RP (38.26). Two values in the plot appear to be smaller than expected, BCUT and NN (3.38) and TI and RP (12.52).

Conclusions

All of the statistical methods were effective in the sense that p-values were small for all types of molecular descriptors. It was surprising that there was not a clear winner either for statistical method or for molecular descriptor. It is not clear which statistical method should have won as each of the methods is highly touted. That the descriptors work at all, given their simplicity or abstractness in the case of BCUTs and TIs, might be considered surprising. The p-values for the three replicates differed by more that we expected; this suggests that cross validation studies need to be replicated (Typically replication of cross validation is not done.) The apparent interaction between the chemical descriptors and the statistical methods suggests that simple validation studies using one type of chemical descriptor and one statistical methods might miss good opportunities. The apparent interaction also suggests that specific descriptors and statistical methods might be exploited to advantage.

References:

Burden, F.R., Winkler, D.A. (2000) A quantitative structure-activity relationships model for the acute toxicity of substituted benzenes to *Tetrahymena pyriformis* using Bayesian-regularized neural networks. Chem. Res. Toxicol. 13, 436-440

ChemTree is described at http://www.goldenhelix.com

Cronin, M. T. D., Gregory, B. W., and Schultz, T. W. (1998) Quantitative Structure-Activity Analyses of Nitrobenzenes to *Tetrahymena pyriformis*. *Chem. Res. Toxicol.* **11**, 902-908.

DEREK is described at http://lhasa.harvard.edu/

DRAGON download from http://www.dist.umimib.com

Gasteiger, J., Zupan, J. (1993) Neural Networks in Chemistry. *Angew. Chem. Int. Ed. Engl. 32*: 503-527.

Hastie, T., Tibshirani, R., Friedman, J. (2001) *The Elements of Statistical Learning* Springer, New York

HAZARDEXPERT is described at http://www.compudrug.hu/hazard.html

JMP is described at http://www.jmpdiscovery.com

Rusinko, A., M. W. Farmen, et al. (1999). Analysis of a large structure/biological activity data set using recursive partitioning. Journal of Chemical Information and Computer Sciences 39, 1017-1026.

Todeschini, R., Consonni, V. (2000) *Handbood of Molecular Descriptors* Eds. Mannhold, Kubinyi and Timmerman. Wiley-VCH Weinheim, Germany.

TOPKAT is described at http://www.accelrys.com/products/topkat/

Venables, W.N., Ripley, B.D. (1999) *Modern Applied Statistics with S* Springer, New York.

Tables

Table 1. Data sets used, their size, web site for description and where they can be obtained.

Data Set	Size	Description	Response Type
NCI	~29,000	HIV	1/0
Yeast	~9,000	PC Score of 6 variables	Continuous
Mutagenicity	~1,800	Ames test	1/0
Toxicity	~250	LC50 aquatic organism	Continuous

These data files can be downloaded from www.niss.org.

Table 2. Descriptor set name and numbers of descriptors. We also give the number of eigenvalues for each set for the NCI data set greater than one to indicate the approximate dimension size of the set.

Descriptor Set	Number of Descriptors	Number of Eigenvalues >1
BCUT	64	4
Constitutional	47	14
Topological	260	30
Other	247	57

Table 3. Lease squares means of Log(P-value) for interaction of statistical method and descriptor type.

	Descriptor Type			
Statistical Method	Cons	Frag	TI	BCUT
PLS	28.17	43.15	18.58	18.67
Neural Network	21.26	19.09	28.45	3.38
Recursive Partitioning	22.76	38.26	12.52	56.95

Table 4. log10(P-value) for fit of observed versus predicted for test set using a model from the training set.

ID	DataSet	Descriptors	StatMethod	-log10(p-value)
1	Mut	Cons	RP	27.00*
2	Mut	Cons	RP	30.28*
3	Yeast	Cons	NN	42.52
4	Tox	Frag	NN	2.76
5	Tox	BCUT	PLS	2.34
6	NCI	TI	RP	1.71
7	Mut	Cons	PLS	34.05
8	Tox	Frag	PLS	26.82
9	Mut	BCUT	NN	9.26
10	NCI	Frag	PLS	36.22*
11	NCI	Frag	PLS	28.45*
12	NCI	BCUT	RP	46.14
13	Yeast	TI	PLS	39.84
14	Mut	TI	NN	29.00*
15	Mut	TI	NN	39.67*
16	Mut	Frag	RP	44.14
17	Yeast	BCUT	PLS	39.93
18	Tox	Frag	RP	21.93

* Replicate pairs.

Table 5. Analysis of Variance

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	14	3614.0438	258.146	8.3732
Error	3	92.4901	30.830	Prob > F
C. Total	17	3706.5339		0.0528

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Data	3	3	973.4229	10.5246	0.0422
Des	3	3	239.9667	2.5945	0.2272
Stat	2	2	256.7875	4.1646	0.1363
Des*Stat	6	6	1672.4314	9.0411	0.0492

Table 6. Lease Squares Means

	Cons	TI	BCUT	Frag
PLS	28.17	18.58	18.67	43.15
NN	21.26	28.45	3.38	19.09
RP	22.76	12.52	56.95	38.26

Figures

Figure 1. Prototypical feed-forward NN with one hidden layer (3 hidden units)

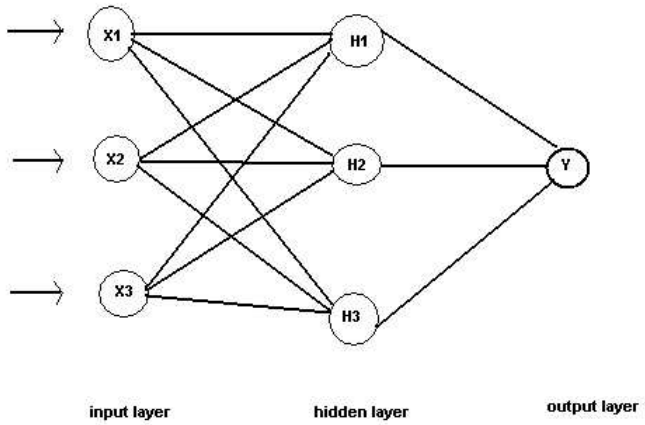


Figure 2. First split for recursive partitioning tree for Tox data set and BCUT descriptors.

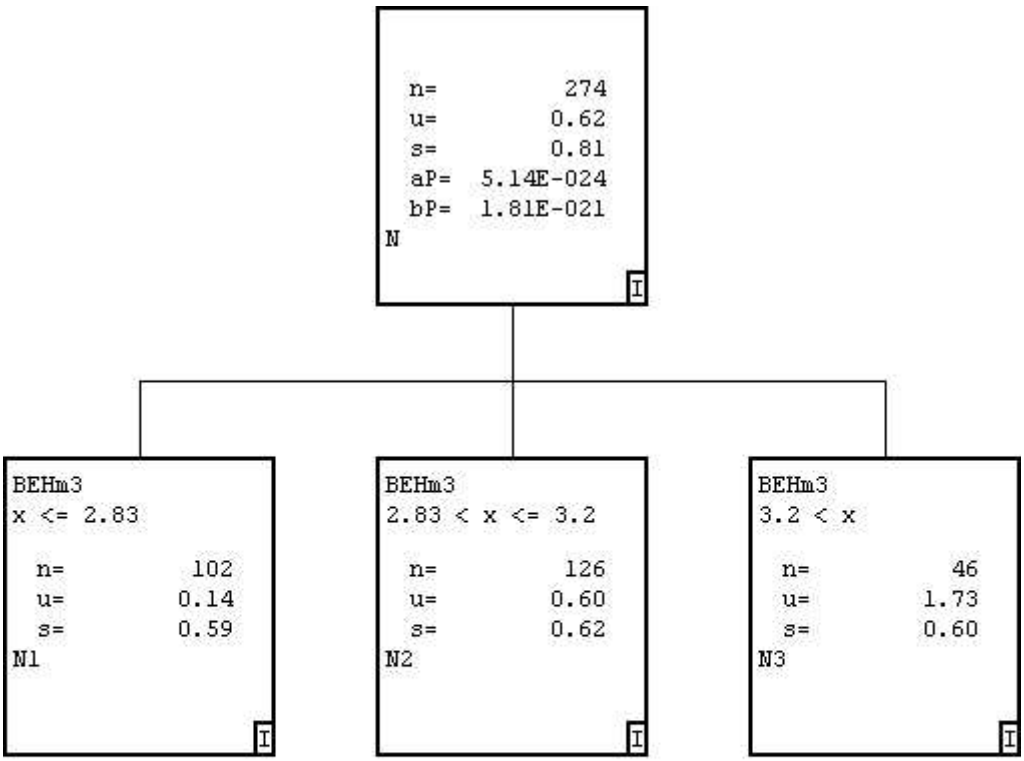


Figure 3. Observed versus Predicted for Tox data set, Frag descriptors and recursive partitioning analysis.

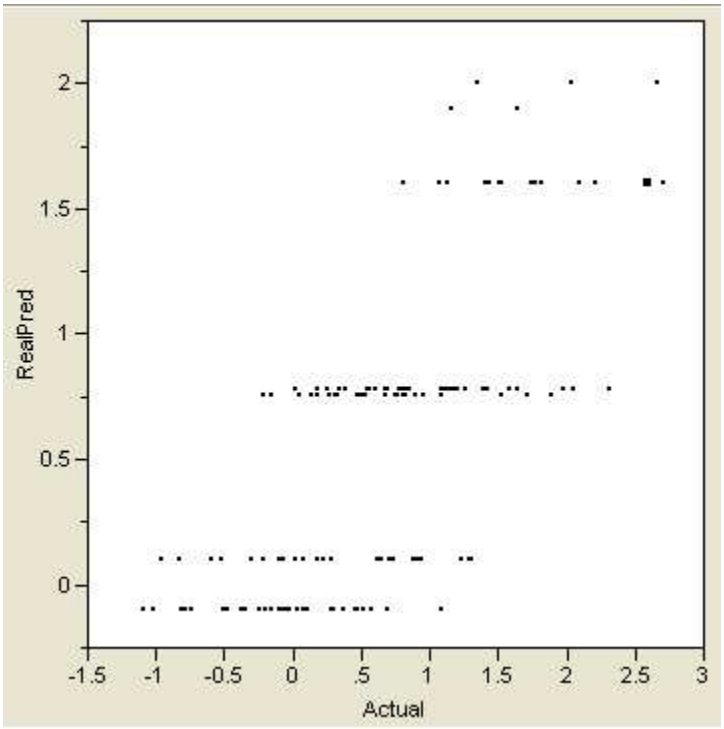


Figure 4. Plot of Least Squares Means versus their rank for interaction of statistical method and chemical descriptor.