SOME STATISTICS IN BIOINFORMATICS

Patty Solomon

School of Mathematical Sciences, The University of Adelaide, Adelaide, SA 5005, Australia

SUMMARY

The spirit and content of the 2007 Armitage Lecture is presented in this paper. To begin, two areas of Peter Armitage’s early work are distinguished: his pioneering research on sequential methods intended for use in medical trials, and the comparison of survival curves. Their influence on much later work is highlighted, and motivate the proposal of several statistical “truths” which are presented in the paper. The illustration of these truths demonstrates biology’s new morphology and its dominance over statistics in this century. An overview of a recent proteomics ovarian cancer study is given as a warning of what can happen when bioinformatics meets epidemiology badly, in particular, when the study design is poor. A statistical bioinformatics success story is outlined, in which gene profiling is helping to identify novel genes and networks involved in mouse embryonic stem cell development. Some concluding thoughts are given.

KEY WORDS: sequential trial; survival analysis; some statistical truths; AIDS epidemiology; microarray data analysis; bioinformatics; serum proteomic test for ovarian cancer; embryonic stem cells; pluripotency; gene expression; gene profiling

1. INTRODUCTION

It is a pleasure to dedicate this paper to Professor Peter Armitage in recognition of his inspired and deep contributions to statistics which now encompass more than 50 years. A contribution to me personally was, in 1985, to act as external examiner of my PhD, which was supervised by David Cox. The spirit and content of this paper are intended to reflect the 2007 Armitage Lecture which I presented in Cambridge on 17th November 2007, with both Peter Armitage and David Cox in the audience. I begin in Section 2 by describing some of Peter Armitage’s best known work on sequential clinical trials, which has its origins in sequential methods for industry applied during World War II, and his influential work on the comparison of survival curves. Peter Armitage’s research demonstrates a truism, that innovation in statistical thinking and methods is best driven by substantive applications, and this provides a basis for me to propose some statistical “truths” in Section 3. After the sequencing of the human genome, biology rapidly transformed from a small-scale laboratory-bench research discipline, to one more often conducted on an industrial scale, and increasingly dependent on the mathematical, computational and physical sciences for advancement. Some part of the role statistics can play in this new mode of biological research is the focus of Section 3. It is also argued in Section 3 that biology is likely to be the dominant application area motivating new statistical development in this century, just as biology, industry and medicine might be said to have done, sequentially, in the last one.

I then discuss two specific examples for illustration. Things can go badly wrong when bioinformatics meets epidemiology if the former is not well-informed about the latter, and an ovarian cancer case-study is presented in Section 4 which illustrates the very real dangers of poor study design. This is the now discredited Petricoin et al. Lancet study [1] which claimed to have derived a simple early-stage cancer diagnostic ‘test’ for ovarian
cancer based on proteomic spectra of blood and a genetic algorithm. A more successful example, a meeting of statistical bioinformatics and gene expression data, is presented in Section 5, where I describe a new method for analysing microarray time course data which follows a pre-defined time course profile [2]. Novel genes involved in mouse embryonic stem cell development are identified and linked to protein interaction networks. There is much scope for adapting existing biostatistical techniques in the bioinformatics context, which are largely unexplored to date, and some ideas are briefly discussed in the concluding section.

2. Peter Armitage’s early work

There are two areas of Peter Armitage’s early research which I believe to be particularly distinguished by their originality and path-breaking contributions to statistical science. The first is his well-known and much admired pioneering research on sequential methods intended for use in medical trials. Much of this work was brought together in *Sequential Medical Trials* (SMT), a book widely regarded as being many, many years ahead of its time [3]. Peter Armitage’s involvement with sequential methods began during World War II, when he was assigned to SR17, a Unit in the UK Government’s Ministry of Supply headed by George Barnard. Peter Armitage talks about this engrossing early start to his career in a Conversation with Vern Farewell and David Spiegelhalter (Armitage Day 2007 [4]). In Unit SR17, Peter Armitage worked with other mathematicians rapidly coming to grips with the complexities of applied statistics on practical applications of quality control, sampling inspection of ordinance, among other things. He recognised the need to move away from infinitely long sampling schemes in sequential methods and rather, to work with curtailed schemes, and Wald’s work was influential at this time. In the interview, he further comments on how very few statistics textbooks there were in the early 1940’s, and how difficult it was to get a grasp of the subject as a whole. This fact, coupled with the realization that statistics can make a really useful contribution to society, no doubt provided strong motivation for Peter Armitage to write his own book on sequential methods and subsequent valuable texts.

At the end of the European War in 1945, Peter Armitage spent a year at the University of Cambridge and worked at the National Physics Laboratory, before taking up a post at the MRC Unit in London. The period of intense industrial war-time activity in sequential methods was then followed by the post-war era of controlled clinical trials, which compared the effectiveness of different therapeutic or prophylactic treatments. Today, sequential methods in medical research often use procedures that are known collectively as repeated significance tests, and Peter Armitage formulated much of his original sequential trials methodology on the comparison of two treatments, $A$ and $B$, and pairs of patients, with one member of each pair being randomly allocated to $A$ and one to $B$. The data are then analysed as the results of each pair became available. The Preface to SMT records that ‘Sequential analysis has an immediate appeal in clinical research’ [3].

SMT, and its much lauded 1975 Second Edition [5], is an absolute a gold-mine of good statistical advice. The principles of controlled trials in general are discussed, and the principles of good experimental design elucidated. Peter Armitage points out that these issues present themselves ‘in almost every field of biological experimentation, for it is the nature of biological material to show some degree of unpredictable variability’ (SMT Section 1.2). Further, he writes, ‘It has been generally recognized, since the work of R.A. Fisher in agricultural research in the 1920’s, that a valid comparison can be achieved only by some form of randomization . . . ’ (SMT Section 1.2) and indeed, therein lies the nature of our discipline. Chapter 1 of SMT also outlines how patients should be allocated either completely at random, or with an element of randomization, to the groups which will receive various treatments, that serial entry may afford an opportunity to use some effective and adequate substitute for restricted randomization, that patients should be entered into the
study *before* the randomization list is consulted, and so on. The pearls of wisdom just keep on coming and the advice is as pertinent today as it was 50 years ago.

A second major and influential piece of Peter Armitage’s early work, on which I would like to focus, is the comparison of survival curves. It was recognised post-war that using a linear model with normal errors to relate a single response variable to a large number of explanatory variables was unlikely to provide an appropriate framework for the analysis of survival data. The response variables in survival data have awkward distributions and are subject to censoring. Only the approach of the actuaries, *i.e.*, life-tables, was well-established and known at this time, but this method could not handle a large number of explanatory variables. Peter Armitage’s seminal Royal Statistical Society read paper [6] examined the asymptotic relative efficiencies of four methods for comparing survival-time distributions when these are exponential and individuals enter the study at a uniform rate during the interval \((0, T)\), the analysis taking place at time \(T\). The exponential distribution was chosen as being an “obvious one to start with” [4], having the attractive property that the maximum likelihood estimate of the hazard of death is the number of deaths divided by the total survival time. The four methods considered were maximum likelihood, the sign method, a comparison of the proportions of survivors at age \(\tau\) (age being measured from entry) using only those individuals entering in \((0, T - \tau)\), and the Kaplan-Meier survival method for the analysis of actuarial data [7]. The sign method was chosen as being particularly suitable for sequential analysis, in which individuals entering are paired and their survival times compared.

Although the exponential distribution has been criticised as a model for human survival, it has in fact found important applications to such. One notable example was in modelling survival following a diagnosis of AIDS in England and Wales in the path-breaking Report of a Working Group (the Cox Report [8]). In this report, Reeves fitted via maximum likelihood an exponential survival distribution to time from diagnosis to death to the first 997 UK AIDS cases reported to the Communicable Disease Surveillance Centre and Communicable Diseases (Scotland) Unit to the end of September 1987. Her model was supplemented by a probability of zero survival time to account for the significant proportion of cases for whom a diagnosis of AIDS and death were virtually simultaneous (*i.e.*, those men who were effectively diagnosed after death). Figure 2.1 shows the estimated survival curve from this model, together with the non-parametric Kaplan-Meier estimates. The estimated constant hazard of death with median survival of 9.22 months observed for the England and Wales cohort was a finding echoed in other western countries around the world in the mid to late 1980s.

Parenthetically, it is worth noting that the Cox Report itself represents an important example of innovative statistical thinking being driven by a substantive application. In addition, this international confluence of expertise from medicine, virology, public health, social science, mathematics and statistics, set a new benchmark for collaborative research efforts for the common good. Here, it was aimed at the prevention of a major national public health disaster (the potential spread of HIV/AIDS through the general UK population) and the timely development of effective treatments for a new and deadly human virus.

David Cox was a discussant of the Armitage survival paper [6], and, of course, subsequently, he himself presented some of the most important work in modern statistics, namely, his semi-parametric proportional hazards regression model [9]. If there was a Nobel Prize for Statistics, then this early work on survival analysis and Cox’s model would surely have won it. Clinical trials for heart disease, stroke and cancer alone have saved and/or prolonged millions of lives, and improved the quality of life for many more. In his discussion of Peter Armitage’s RSS read paper, Professor Austin Bradford Hill wrote “I am glad that Dr Armitage has turned his attention so helpfully to a problem that can be very troublesome in clinical medicine” [10]. Peter Armitage states in the Introduction to the paper that his work was motivated primarily by the possibility of using sequential methods for the design and analysis of clinical trials for the treatment of chronic diseases, in which the main outcome of interest was length of survival following treatment. This motivation
FIGURE 2.1. Kaplan-Meier estimates and estimated exponential probability of survival with constant hazard of death following a diagnosis of AIDS for homosexual men in England and Wales to September 1987 (Cox Report [8]). The exponential model is adjusted for a probability of zero survival time for those cases for whom a diagnosis of AIDS was made at or after death. Survival time is measured in months since diagnosis.

for substantial theoretical development by an important (life-or-death) problem leads me to some “truths”, which I hope are universally acknowledged, implicitly if not explicitly, in the statistical world.

3. SOME TRUTHS

There are five “truths” that I should like to identify.

I: Innovation in statistical thinking and methods is best driven by substantive applications.

II: Biology is dominating statistics at the beginning of this century, just as it did at the beginning of the last one.

III: Statistics is a fundamental, enabling discipline.

IV: Innovation in bioinformatics and systems biology depends critically on high-level interactions between biologists and researchers from cognate disciplines.

V: Statistical Science must itself be strong to enable high-level collaborations with scientists from cognate disciplines.

I will deal with these in order but at varying length.

Truth I: Innovation in statistical thinking and methods is best driven by substantive applications.

An early example of my first “truth” is the statistical analysis of noisy space probe images in the 1950s; nowadays, subtracting ‘graininess’ from raw digital images is a standard tool in many fields. Similarly, the demand for financial services, where the main objective is to reduce exposure to risk, has driven innovation in time series analysis, stochastic analysis and chaos theory.
Sadly, innovation in statistical thinking and methods can often be driven by disasters, and we have already mentioned the AIDS pandemic. Amongst the many early statistical challenges presented by AIDS was that of conducting placebo-controlled drugs trials, when randomizing patients to the control arm meant certain death. In desperation, participating patients got together to pool their tablets to share them out equally, so ensuring that everyone received some active drug. Some patients enrolled in a multicentre trial who believed they had been randomized to placebo dropped out and attempted to re-enrol at another centre. One flexible approach to trial design introduced early on was the use of subtrials. In the AIDS trial ‘Alpha’, patients intolerant of AZT were offered a choice of subtrials: patients choosing subtrial A were randomized onto high DDI (didanosine) versus low DDI versus placebo; patients choosing subtrial B were randomized onto high DDI versus low DDI. However, the randomization scheme did not work very well, in that very few patients opted for the subtrial containing the placebo group and this arm of the trial was wound up early (private communication, Peter Armitage). The emphasis in rethinking trial design in this way was on making such trials flexible and acceptable to potential participants. The subtrial idea also came in to the cardiovascular trial ISIS-3 (International Study of Infarct Survival, 1992), in which patients for whom physicians thought fibrinolytic therapy was required were randomized to one of three active treatments, whereas those for whom the indication was ‘uncertain’ were randomized to one of these three groups or to ‘open control’, but no placebo [11]. Thus the challenges presented by the early AIDS trials led to valuable contributions to trial design, as well as to methods for handling dropouts, noncompliance, and missing data [12].

Backcalculation, based on a convolution equation represented as a simple susceptible, infected and infectious, removed ($SIR$) epidemic model, was developed for reconstructing the time-varying HIV infection incidence and predicting future cases of AIDS (see [13, 14, 15, 16]). Backcalculation was not without controversy in its application, but there can be no doubt it was valuable in providing health policy makers both with an understanding of the natural progression of HIV disease, and in providing quantitative working estimates of the (then) unknown extent of the unobserved HIV infected population.

A great deal of other statistical epidemiological work was also stimulated by AIDS related research, another notable example being the staged Markov models based on CD4$^+$ T-lymphocytes for the natural history of HIV infection. Early work on this was by Longini and co-workers [17].

Truth II: Biology is dominating statistics at the beginning of this century, just as it did at the beginning of the last one.

My second “truth” is supported by recent statistical work on biologically based problems, recalling that early 20th Century work in statistics, e.g., that of Pearson and Fisher, also arose from biological investigations. For example, in recent years, image analysis for microarray experiments has attracted much attention. In microarray informatics, research questions are still being addressed about when and how to correct for background noise, where there may be a significant variance-bias trade-off problem (see for instance, Silver et al. [18]). No background correction usually leads to the presence of (substantial) systematic experimental bias, but over-correcting for background may induce too much variability into the data, especially in the low-abundance signals which are usually the ones of primary interest to biologists. Figure 3.1 shows an image of a hybridized two-colour human long-oligonucleotide array with 20,000 spots on the slide. Nineteen thousand of these represent sequences of oligonucleotides between 100 to 300 bases in length and correspond to mostly unique genes; the remainder are controls of various types. The slide is from a series of control experiments conducted at the Adelaide Microarray Facility in which a mixture of mRNA from MCF7, a breast cancer cell line, and Jurkat, a T-cell line, were mixed and applied to the array. These cells have quite different patterns of gene expression and produce lots of green and red spots, as can be seen in Figure 3.1. Yellow spots indicate equal hybridization for the two mRNA samples for the sequence. The Figure shows a relatively good image of a long-oligonucleotide array, but it still has a number of
Figure 3.1. An image of a hybridized human two-colour long-oligonucleotide microarray from a control experiment in which mRNA from MCF7 (a breast cancer cell line) and Jurkat (a T-cell line) were hybridized to the array. There are 20,000 spots on the slide, with 19,000 of the long oligonucleotides corresponding to mostly unique genes. Most positive controls are cDNA and are visible as bright diagonal lines. This is a good image of an oligonucleotide slide, but it still has a number of minor problems from an image processing point-of-view: there is a vertical scratch starting near the top; a print of a small bubble; there is some high background near the bottom of the slide which shows as ‘black’ spots; and too many spots are white owing to the scanner settings being too high.

minor problems with respect to processing the image to extract the quantitative expression data for analysis; these are outlined in the caption to the Figure. Note that the image as presented in Figure 3.1 represents the raw data obtained from a single-slide microarray experiment.

Figure 3.2 shows the ‘rawest’ of raw data from an Affymetrix GeneChip array produced as part of a grape-berry development project conducted in Adelaide in collaboration with researchers from CSIRO. In the full time-course experiment, the development of berries on the grape-vine was observed at different geographical sites under different environment and treatment conditions over a two-year period; Figure 3.2 shows just one chip from this experiment. Affymetrix technology is different to that for two-colour microarrays: an Affymetrix chip has hundreds of thousands of short oligonucleotide probes laid out in sets representing up to 40,000 target sequences or genes; it involves the hybridization of only one mRNA sample to the chip; and has a more formalized experimental protocol and procedure. Due to the increased accuracy and precision of the technology, the main image analysis issue in processing Affymetrix chips is checking the grid alignment.

The sequencing of the human genome, together with the development of increasingly accurate high throughput technologies, has led to the mathematization of biology which is now much more quantitative than in the past. In the beginning of microarray data analysis, we clustered or looked for differentially expressed genes using a statistic, such as $t$ or $F$, and produced lists of ranked genes based on suitably chosen cut-off values [19]. Later, we examined these clusters or lists of differentially expressed genes for enrichment with various pre-defined sets of genes, such as the Gene Ontology categories. Later still, we skipped the cut-offs and the lists, and kept the statistics and the pre-defined sets of genes, and looked among them for the differentially expressed gene sets using the statistic on all
FIGURE 3.2. An image of a processed Affymetrix GeneChip array from an experiment on grape-berry development conducted in Adelaide, Australia. This image is the ‘raw’ data obtained from the .CEL file stage in the processing of the chip. Roughly 500,000 25-mer probes are laid out on the chip in probe sets, representing 40,000 target sequences or genes.

The genes. Both types of gene-set analysis use just the pre-defined sets of genes, but do not make use of the relationships or interactions among the genes within the sets. The genes within each pre-defined set are all treated equally, and the sets are pre-defined, not discovered from the data. Gene networks took us one step further in the evolving sequence of methods for the analysis of gene expression microarray data, where we try to make use of the relationships (i.e., interactions) between the genes within the sets, as illustrated in Figure 3.3. The left-hand diagram in the Figure gives a schematic of the molecular interactions involved in the Citrate Cycle (i.e., the tricarboxylic acid cycle) in which the fuel molecules of fats, sugars and amino acids are oxidized (i.e., burned) to produce carbon dioxide. The yeast interaction networks shown on the right of Figure 3.3 represent some of the most widely studied systems in genomics and bioinformatics, and studies on *Saccharomyces cerevisiae* gene networks are helping to shed light on the pathways involved in the more complex genetics of other organisms. Both the yeast metabolic interaction network and the protein-protein interaction network exhibit dense local neighbourhoods, suggesting for the latter that the position of a protein in part of a network may predict interactions with other proteins or genes. As well as providing insight into gene-interactions, the metabolic interaction network can help determine the genetic and regulatory pathways which underlie the growth and development processes.

We are now in the post-genome era of bioinformatics and systems biology. We can think of bioinformatics as encompassing all quantitative work at the interface of the biomedical, physical, computational and mathematical sciences, in the pursuit of understanding the genetic basis of disease and related phenomena. Systems biology can be defined as ‘the study of an organism, viewed as an integrated and interacting network of genes, proteins and biochemical reactions, which give rise to life’.\(^1\) It is highly unlikely that other application areas, including those of finance and risk, have the same potential as biology to attract statisticians and dominate statistical development.

*Truth III: Statistics is a fundamental, enabling discipline.*

Statistics as a discipline has its own internal dynamics and coherence. But we are also experts at dealing with uncertainty and variability, and have skills in experimental design,

\(^1\) Systems Biology - the 21st Century Science, Institute for Systems Biology 2008

http://www.systemsbiology.org
Yeast protein-protein interactions
Yeast metabolic interactions
Some molecular interactions in the genome

Figure 3.3. Illustrations of molecular interactions and networks (private communication, T.P Speed). Yeast (Saccharomyces cerevisiae) is one of the most widely studied organisms in genomics and bioinformatics. It provides canonical models for a number of systems, such as the time course experiments associated with the yeast cell cycle. Yeast gene interaction networks can shed light on the more complex genetics and inherited phenotypes in other organisms. The yeast metabolic interaction network and protein-protein interaction network shown on the right of the Figure exhibit dense local neighbourhoods, the latter suggesting, for example, that the position of a protein in the network may be predictive of other protein interactions. The metabolic interactions shed light on the genetic pathways and interactions underlying the growth process. The picture of the left is a schematic of the genetics and molecular interactions associated with the Citrate Cycle. Also known as the TCA cycle (i.e., the tricarboxylic acid cycle), this is an oxidation cycle in which fuel (fats, sugars and amino acids) molecules are oxidized to form carbon dioxide.

data analysis, reasoning and synthesis. This is why we can move so readily into new areas such as bioinformatics and systems biology. We need to do so because good statistical analysis is the key to getting the best out of these new technologies, and there is a critical shortage internationally of quantitatively skilled biological graduates. Because we are almost always aiming to measure low abundance transcripts or signals in the presence of high systematic (i.e., experimental and laboratory) and biological variability, we are constantly put at the limits of our analytical capabilities. This challenges us and drives novel development. 

Truth IV: Innovation in bioinformatics and systems biology depends critically on high-level interactions between biologists and researchers from cognate disciplines.

Biological research has a new morphology; it is no longer a data-poor, small-scale, laboratory-based research discipline, but one that is data-mega-rich with sources of systematic and random error in its experiments akin to those from a large-scale industrial
process. As a result, biological concepts, models and practice are now more quantitative and more critically dependent on the expertise and perspectives of other scientific disciplines; in particular, the physical, computational and mathematical sciences. I believe that advancement in knowledge in biomedical bioinformatics is unlikely in the absence of high-level multidisciplinary collaboration amongst experts from the cognate disciplines.

*Truth V: Statistical Science must itself be strong to enable high-level collaborations with scientists from cognate disciplines.*

This is a simple truth that in order to be good and effective collaborators, statisticians must bring their own unique strengths to any collaboration. These will derive from training in the principles and practice of statistics, an awareness of developments in statistical methodology, and having the time and resources to develop new methods, in the context of the substantive applications area, as required. Bioinformaticians in particular may need time to acquire the appropriate knowledge in molecular biology and new biotechnologies if they are to be effective collaborators with biologists. Although Truth V may seem a simple and obvious truth, it is not necessarily widely acknowledged. For example, in my own home country of Australia, chronic underfunding of the mathematical and statistical sciences by successive federal governments has led to a long-term critical shortage of statistical graduates. Consequently, there is now a critical lack of suitably qualified biostatisticians and bioinformaticians, both areas of national, and international, need.

Finally, we should have seen the current situation coming (some of us did), and were warned with what could be an additional “truth”.

*Truth VI: “When we entered the era of high technology, we entered the era of mathematical technology”* (Ad Hoc Committee on Resources for the Mathematics Sciences, US National Research Council, 1981).

I now turn to my two examples.

4. **Statistical Bioinformatics meets Epidemiology**

Simply put, DNA makes RNA makes protein. Whereas microarrays allow us to measure the mRNA complement of a set of cells, mass spectrometry allows us to measure the protein complement (or subset thereof) of a set of cells. Proteomic spectra are mass spectrometry traces of biological specimens and an example of a mass spectrum of human serum is shown in Figure 4.1. There are numerous technologies, including MALDI-TOF (matrix-assisted laser desorption and ionisation-time-of-flight) in which pulses of laser light provide energy to ionise and vaporise the protein sample, SELDI-TOF (surface-enhanced laser desorption and ionisation) which is a variation of MALDI in which the sample-matrix is spotted onto a slide, and Qq-TOF which increases the mass resolution achievable by other TOF mass spectrometers. The mass-to-charge ratio (m/z) shown in the top plot of Figure 4.1 is determined from the length of the tube, the kinetic energy given to the ions by the electric field and the time of flight.

There is a great deal of interest in discovering protein biomarkers in blood to diagnose cancers early on. Internal cancers such as ovarian and prostate cancer are particularly difficult to diagnose early, and this is why the application of proteomics mass spectrometry in this area has generated so much excitement. Protein profiles are being assessed using serum and urine from patients, rather than by invasive tissue biopsies. Moreover, proteomic spectra are cheaper to run on a per unit basis than microarrays, and samples can be run on large numbers of patients.

An (in)famous early study by Petricoin, Ardekani and others from the USA’s Food and Drug Administration’s own proteomics group [1] used proteomic patterns in serum to identify ovarian cancer at an early stage. Ovarian cancer is frequently a deadly disease, and its morbidity and high mortality rate are strongly linked to the inability to detect the tumours at an early stage. A simple, easily applied diagnostic test with high sensitivity and specificity would be of great utility. The Petricoin study involved 100 ovarian cancer patients, 100 normal control patients, and 16 patients with ‘benign disease’. Proteomic spectra from
What Do the Data Look Like?

Figure 4.1. A mass spectrum of human serum. The mass-to-charge ($m/z$) ratio for each ion is determined from the length of the tube, the kinetic energy given to the ions by the electric field (see main text for discussion) and the time of flight. This allows the spectrum of the intensity of the ions against the $m/z$ value to be plotted (top spectrum). The intensity of the ions may also be plotted against the time of flight (bottom spectrum).

50 cancer patients and 50 normal controls were used to train a Genetic classification algorithm, which was then tested on the remaining spectra. The results were spectacular: the classifier correctly classified 50 out of the 50 ovarian cancer cases, correctly classified 46 out of 50 of the normal controls, and correctly classified 16 out of 16 of the benign disease as ‘other’.

There was much excitement in response to the publication of the Petricoin Lancet paper, and research groups around the world started asking how they could do this type of analysis with their types of cancer. Almost immediately however, experts in cancer research and clinical chemistry began to raise questions about the approach (Eleftherios Diamandis of the Mount Sinai Hospital, University of Toronto, for one) saying in large part that it really should not work. This is essentially due to limitations of the technology, which is that mass spectrometry has a rather limited dynamic range, so that it cannot see trace elements at the same time as abundant ones. Proteins shed into serum from ovarian tumors while the tumors are still small are likely to be present in only trace amounts, and most profiles of serum notice blood proteins and miss the smaller ones. There are several things that are detected which are real, but most of these are acute phase reactants (or fragments thereof) which tend to indicate that the patient is rather sick, but are not very specific as to the disease (see Zhang et al. [20] which encounters some of the limitations of the technology). Various questions about oddities on the data began to crop up at this time as well. Sorace and Zhan [21] point out some of the problems, as do Keith Baggerly and colleagues from the M.D. Anderson Cancer Center (Houston, Texas) who could not reproduce the published results from the available data [22] and who found that the ‘perfect’ classification of peaks
Figure 4.2. From Figure 6A of Conrads et al. [24], record count by run date for the higher quality instrument Qstar. Record count is the total number of data points and is used as a quality control measure. Colour represents day: red circles = day 1; green squares = day 2; blue triangles = day 3. Quality control has deteriorated by day 3. Reproduced by Baggerly [25].

was achieved entirely within the ‘noise’ region of the data [23]. Baggerly et al. [22] detail numerous data and analytical issues of concern, including that there was an apparent change of protocol near the end of the dataset, that there was no time-$m/z$ calibration done (which is a known source of bias), and that there was no evidence of randomized order of processing. Furthermore, there is nothing in the paper about the epidemiology of the study, that is, about how the samples were collected or processed, or any demographic or clinical information about the patients apart from their case-control status. All this strongly suggests a qualitative difference in how the samples were processed, and possibly not just a difference in the biology.

Not to be deterred, in January 2004 (the same month that Baggerly et al.’s Bioinformatics paper appeared online) companies Correlogic, Questdiagnostics and Labcorp announced plans to offer a test called ‘Ovacheck’ for availability by mid-year for which samples would be sent by clinicians for diagnosis. The estimated market for this “home-brew” test was 8 to 10 million women at an estimated cost of $US100-$200 per test. In February 2004, the New York Times covered the story, noting some potential problems, and the Society for Gynecologic Oncologists released a position statement saying that a test seemed premature. The FDA sent letters to the companies preparing to market the test, asking them to withhold it. In the meantime, the Petricoin group published what turned out to be an abortive follow-up paper [24]. In this second study, Conrads et al processed samples with their original SELDI technology and also with a higher resolution instrument Qstar (Qq-TOF), and added some quality assurance/quality control steps to remove bad spectra. They still used patterns for analysis, and the reported results were even better than in the original Lancet study. They demonstrated 100% sensitivity and 100% specificity for identifying cancer from normal, and stated that this “emerging paradigm” is ready to go to a large clinical study. So what was going on?

Figure 4.2 is taken from Conrads et al. [24] and shows the record count by day for the higher-accuracy instrument Qstar. The record count is the total number of data points.
Figure 4.3. From Figure 7 of Conrads et al. [24], record count by run date, and the time order controls were processed in black (upper plot) and the time order cases were processed in black (lower plot) superimposed. Record count is the total number of data points observed and is used as a measure of quality control. Colour represents day: red circles = day 1; green squares = day 2; blue triangles = day 3. Clearly the processing of cases and controls was not randomized. Reproduced by Baggerly [25].

observed and is used as a quality assurance measure. The three days are represented by different colours and symbols: the day 1 counts are red circles; the day 2 counts are green squares; and the day 3 counts are blue triangles. Clearly quality is deteriorating as time goes on, and is poor by day 3. They do something about it, but we also learn in passing that the controls were mainly done on day 1 and some at the start of day 2, whereas the cancers were processed on days 2 and 3, in a context where something (i.e., quality) was changing. This is shown in Figure 4.3 where the results for the controls and cases are superimposed in order of processing by day on the upper and lower plots, respectively.

Given the changing quality of the data over time, this confounding of case-control processing with changing quality control biases the results, as the cancer samples were more affected by the worsening problem on day 3. Again, as for the initial Petricoin study, this is before one even gets to the epidemiology of the study, or the sample preparation. Obviously, there is no way a woman should be told she needs an oophorectomy based on these tests. Finally, in June 2004, the FDA ruled that ‘Ovacheck’ could not be made available as advertised under the “home brew” exemption, as the software program was a ‘device’ that needed to be more tightly regulated. In September 2006, the FDA released draft guidance on ‘In vitro diagnostic multivariate index assays’ (IVDMIAS), and these rules are being debated even now.

The moral of the story is that a better machine (the Qstar instrument) will not save you if the study design is poor. Even a passing familiarity with the relevant sections on study design from the authoritative Statistical Methods in Medical Research [26], now in its fourth edition, could have saved Petricoin and colleagues from this breakdown in scientific method.
5. GENE PROFILING FOR A TIME COURSE MICROARRAY EXPERIMENT IN STEM CELLS

I now describe briefly a new method for analysing time course gene expression data developed with colleagues at the University of Adelaide, Jonathan Tuke and Gary Glonek [2]. This statistical work evolved from a collaborative microarray project with researchers from the Rathjen Laboratory at the University of Adelaide, who have developed a mouse embryonic stem cell line to study how pluripotency (stemness) is controlled in murine ES cells. Pluripotency refers to the potential of a cell to develop into more than one type of mature cell, depending on environment, and is an important area of research for such diverse medical areas as organ transplants, the treatment of diabetes and the treatment of spinal injuries. Figure 5.1 presents an overview of the biological system under study. The ability to differentiate into any body cell is present in mice stem cells up to and including day 3. After this, the stem cells become multipotent: they still have the ability to differentiate into different types of cells, but now a limited number. At day 6, it is believed the cells have lost pluripotence and represent what is known as a definitive ectoderm state. By day 9, the cells are committed multipotent cells. The genes referred to in Figure 5.1, such as Oct4 and Sox1, are known to be associated with pluripotency and are included here to demonstrate the biological changes of interest. Oct4, for example, is a gene known to be highly expressed at days 0 and 3 when the cell is pluripotent, but not later in the multipotent state. Our analysis of the experimental data (see Section 5.2) confirms this.

The initial aims of this ‘omnibus experiment’ were to study loss of pluripotency over time, identification of genes specific to each pluripotent state (as commensurate with day), and to rank the genes according to their association with pluripotency. I have called this
Figure 5.2. Design of the stem cell microarray time course experiment. Each arrow represents two hybridizations of the same comparison of mRNA samples from two time points conducted as a dye-swap pair; the arrow head points towards the mRNA sample labelled with Cy5 (red), and the arrow tail towards the sample labelled with Cy3 (green). The parameter $\mu_j$ is the true mean gene expression level at day $j = 0, 3, 6, 9$.

an ‘omnibus experiment’ because the Rathjen team used this single, albeit complex, time course microarray experiment to obtain a set of data which enabled them to address a number of scientific hypotheses of interest. The differing scientific claims on the experiment, itself subject to resource constraints of time and money (only 20 slides were available for hybridization), complicated its design, which is shown in Figure 5.2. Hybridizations were performed for each of the pairwise comparisons between days 0, 3, 6 and 9, except for day 6 versus day 0. There was some uncertainty at the design stage whether cells had lost pluripotency by day 6, and I was the only member of the team who urged that this comparison should be included in the experiment (for obvious scientific and statistical efficiency reasons) but ultimately it wasn’t! There were 16 mRNA samples taken, 4 samples of stem cells harvested on each day, with the 4 replicate cultures obtained from different passages. A passage is a cycle of growth and re-plating of cells isolated from the early embryo, and for the purposes of this analysis we treated these as independent biological replicates. Each arrow in Figure 5.2 represents two hybridizations, with the arrow head pointing towards the mRNA sample labelled with Cy5 (red), and the arrow tail with Cy3 (green). Dye swaps were balanced within each comparison and for each replicate culture. The parameters $\mu$ represent the true absolute mean gene expression levels on each day. Although the design is a compromise between resource constraints and the demands of experimenters, it is close to optimal according to recently developed optimality criteria [27]. The hybridizations were performed at the Adelaide Microarray Facility using the CompuGen Mouse 22K Long Oligo Library (5 comparisons within each stem cell sample). Slides were scanned using Spot, and subsequent analysis was performed in R utilizing the Bioconductor suite of packages.

The hypothetical profile for stemness of interest in this study is shown in Figure 5.3. As pluripotency is restricted to the early stem cells (day 3 or earlier) genes that have high expression levels in cells up to day 3, but low or monotonically decreasing expression levels thereafter, are likely to be associated with the biochemical pathways involved in the pluripotency ability of these cells. An aspect of the analysis of especial interest was in finding those genes in the data which satisfy the (hypothetical) expression criteria over
time for a pluripotent gene. We pursue this in the belief that genes with similar developmental temporal expression profiles may well be involved in similar biological processes (here, pluripotency). We also hope that genes identified in this way will share common sequence motifs in their regulatory region. One existing approach which attempts to do this is the Pareto-front method of multi-criterion optimization, in which a competing set of functions is chosen, each of which measures the association of a gene to a pre-specified profile (Fleury et al. [28], Hero and Fleury [29]). Genes found to be Pareto optimal with respect to these criteria are identified as matching the pre-specified profile. Its main disadvantage is that some genes will be detected as Pareto optimal genes whilst only matching a subset of the pre-specified criteria. In other unrelated work, Ingrid Lönnstedt et al. [30] describe an empirical method for ranking genes based on the inner product between the vector of observed log ratios and a vector of constants which define the profile. This can work well for some profiles, but did not provide useful outcomes in our data. Moreover, there is no standardization for gene-specific variances, so a large inner product is not a necessary and sufficient condition for close concordance. Both of these methods are based on contrasts of the data and neither method worked effectively for our data. We found that by simultaneously testing for all criteria however, gene profiling effectively filters out and excludes genes that are only partially consistent with the required profile (i.e., has greater specificity than previous methods). For each gene, we treat the vector of true gene expression levels as a linear combination of linearly independent vectors chosen to represent the pre-specified time profile. The model is fitted to the data by least squares and the genes are ranked according to a suitable test statistic (a novelty is that we use the Intersection-Union Test, see below) accommodating both hypotheses of equivalence of gene expression and hypotheses of differential gene expression simultaneously.
5.1. The gene profiling model. The true mean gene expression level on day $i$ is $\mu_i$, $i = 1, \ldots, 4$, as shown in the hypothetical expression profile of interest in Figure 5.3, as a function of day. The vector $\mu$ is a vector in $\mathbb{R}^4$, and can therefore be expressed as a linear combination of four linearly independent vectors. We need to choose vectors which represent the criteria for pluripotency represented in Figure 5.3, ensuring consistency in the scale of interpretation of the pluripotency parameters, which we call $\gamma$. In the present example, this corresponds to the choice of

$$
\mu = \gamma_0 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \gamma_1 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \gamma_2 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \gamma_3 \begin{pmatrix} 1/2 \\ -1/2 \\ 0 \\ 0 \end{pmatrix}
$$

which leads to the set of model parameters $\gamma_0 = \mu_9$, $\gamma_1 = \frac{\mu_0 + \mu_3}{2} - \mu_6 > 0$, $\gamma_2 = \mu_6 - \mu_9 > 0$ and $\gamma_3 = \mu_0 - \mu_3 = 0$. These parameters satisfy the pre-specified pluripotency profile described by Figure 5.3. Note that $\gamma_0$ is unconstrained, and could be zero. Inference for $\gamma_1 > 0$ and $\gamma_2 > 0$ can be determined in the standard Neyman-Pearson way; we may also adjust the null value to be some non-zero quantity to make the hypothesis more specific, which we do for our analysis (see Section 5.2).

The equivalence hypothesis test for $\gamma_3$ requires an active demonstration that $\gamma_3 = 0$, or something close to it, and not simply a failure to demonstrate differential expression. Typically, if $X$ is a random vector whose probability distribution depends on a real-valued parameter $\gamma$, then to test whether $\gamma$ is equivalent to zero, a neighbourhood around zero is constructed with the alternative hypothesis of interest that $\gamma$ lies within this neighbourhood, and the null we look for evidence against is that $\gamma$ lies outside the neighbourhood, here in either direction (see, for example, the book by Wellek [31]). The neighbourhood defined by a small-valued parameter $\epsilon$ say, is the maximum that the parameter can vary and still be considered equivalent to zero. The simplest and most common way to test the hypothesis is via confidence interval inclusion (CII). We calculate a confidence interval from the observed data, $(L_\alpha(X), U_\alpha(X))$ where $L_\alpha(X)$ and $U_\alpha(X)$ are random variables such that

$$
P(\gamma \in (L_\alpha(X), \infty)) = P(\gamma \in (-\infty, U_\alpha(X))) = 1 - \alpha
$$

For an $\alpha$-level test, we reject the null hypothesis in favour of equivalence if and only if the interval is contained entirely within $(-\epsilon, \epsilon)$. Note that we can also use CII to test separately whether $\gamma_1$ and $\gamma_2$ are significantly positive. But there is a snag. In an ideal world, $\epsilon$ would be chosen to the largest a gene’s mean log ratio (or the difference between two log ratios) can vary about zero and not be differentially expressed, according to biologists. This is analogous to the ‘minimum clinically significant difference‘ in a superiority trial, for example. So in practice, a working, quantitative understanding of equivalence in gene expression should be decided upon in advance in consultation with biological scientists. Unfortunately, it remains the case that rather little is known about gene-specific variation per se, or their interactions and relationships with other genes, and there are many thousands of genes to consider simultaneously. Such knowledge is of course crucial to deciding upon the null and alternative hypotheses values of $\epsilon$ on a gene-by-gene basis, and although the requisite information and knowledge are gradually accruing over time, as microarrays and other genomics technologies are more widely applied by molecular biologists, it is not available yet. What should we do in the meantime? We do what we always do when faced with crucial missing data; that is, make a sensible choice(s) for $\epsilon$ that works, and examine the sensitivity to that choice.
5.2. The Intersection-Union Test and gene profiling. We use the Intersection-Union test (I-UT) for each gene, on which to base gene-selection, where the null hypothesis is expressed as a union [32]:

\[ H_0 : (\gamma_1 \leq 0) \cup (\gamma_2 \leq 0) \cup (|\gamma_3| \geq \epsilon), \quad \epsilon > 0 \]
\[ H_A : (\gamma_1 > 0) \cap (\gamma_2 > 0) \cap (|\gamma_3| < \epsilon) \]

This is an \( \alpha \)-level test, where \( \alpha = \sup \alpha_i \gamma_i \), and is uniformly most powerful with such composite hypotheses. Our main aim is to rank the genes according to their match with the pluripotent profile and we can modify the method to give a quantitative measure of how closely each gene matches the desired profile. For each gene and each parameter, CII is used to test the associated null hypothesis, and rather than using a fixed significance level, we find the smallest \( \alpha_i \) for each \( \gamma_i, \ i = 1, 2, 3 \), such that null hypothesis is rejected. Then \( \sup \alpha_i \) is used as the test statistic to rank the genes. In fact, for our experiment, rather than calculate \( \alpha_i \) for each \( \gamma_i \), we used the width of the largest confidence interval \( U_i \) for each \( \gamma_i \) that was contained within the rejection region. The infimum \( U \) of the \( U_i \) was then used to rank the genes. If one of the intervals did not lie within the rejection region for a gene, it was excluded from the ranking. We initially took \( \epsilon \) to be unity. In addition, the parameter values were changed to \( \gamma_2 > 1.5 \), to ensure a large difference between the gene expression levels on days 0 and 3. In essence, the distance to the nearest boundary of the rejection region is calculated in terms of standard errors of the estimate, where larger values are indicative of pluripotency.

Of the 22,000 genes on the array, 15 were selected by gene profiling as being statistically significantly associated with pluripotency according to the pre-specified profile, and these 15 genes are shown in Figure 5.4. The top-ranked gene, Oct4, is a transcription factor well-known to be associated with pluripotency. Other genes we know about are Utf1 and Nanog (ranked second) which is associated with undifferentiated embryonic cell transcription [33] and Nanog (ranked 11th) which is central to embryonic stem cell pluripotency [34, 35]. In a recent Nature paper, Wang et al. [36] isolated proteins associated with the protein Nanog, and thus with pluripotency. Wang et al’s protein interaction network is shown in Figure 5.5, and illustrates how Nanog functions in concert with Oct4 and other transcription factors such as Sox2. Sox2 has a quite different expression profile to that depicted in Figure 5.3 however. Sox2 has higher expression on day 0 compared to days 6 and 9, equivalent expression for days 6 and 9, with the gene expression level for day 3 lying in between that of days 0 and 6 to 9. The 10 statistically significant ranked genes following this time course profile are shown in Figure 5.6. Sox2 was ranked at position 4. Although retSDR3 has the desired form, with the largest apparent magnitude, it is only ranked 9th in the set of 10 genes matching the Sox2 profile. The low ranking results from the large gene expression variance of this clone (0.181) relative to the other ranked genes (which have an average variance of 0.054). This illustrates that for two genes with the same coefficient magnitudes, gene profiling will rank lower that gene which has more uncertainty (i.e., higher variance) in its true expression profile.

Note that in each application of gene profiling, novel genes have been identified. Finally, to investigate the potential effects of altering the equivalence neighbourhood defined by \( \epsilon \), we repeated the original Oct4 pluripotent profile analysis for increasing values of \( \epsilon \). We found there is a tendency for more genes to lie within the rejection region as we increase the width of the equivalence neighbourhood \( \epsilon \), as we would expect, but some increased variability in expression levels between days 0 and 3 (which correspond to our equivalence hypothesis days). Gene profiling is demonstrated to be reasonable here: Oct 4 remains the top-ranked gene for \( \epsilon = 0.5, 1.0 \) and 1.5, and even for \( \epsilon = 2 \), which is huge, it has only dropped to rank second.

We have made no direct adjustment for multiple testing here. Fortunately, no such substantive correction was required with only 15 statistically significant genes obtained in the final ranked set for pluripotency. In other applications, with potentially large ranked sets to consider, some control for multiple testing will be necessary. Our research in progress on...
controlling for multiple testing in gene profiling shows that the posterior probability of being ranked (or differential expression, or equivalence, the definition depending on context) provides a valuable method of discriminating the ‘strength of evidence’ in the data. The usual methods of adjusting for multiple testing in bioinformatics based on \( P \)-values and the false discovery rate are not valid for equivalence testing owing to the discontinuity under the null hypothesis. This is consistent with an earlier finding which established the superior of posterior probabilities as a measure of strength of evidence in gene expression data more generally [37].

6. SOME CONCLUDING COMMENTS

Gene profiling is a valuable method for analysing time course gene expression data (although it is by no means limited in application to such) since it provides a statistically valid framework for pattern searching that goes beyond unsupervised learning methods such as clustering. Gene profiling is also proving useful as a tool for exploring regulatory networks and pathways, although care is needed to avoid false positives in this situation, and for studying the behaviour of a system under study based on experiment structure. It is, of course, only one of many novel statistical developments emanating from biology’s new
Figure 5.5. Protein-protein interaction network associated with the protein Nanog, from Wang et al [36]. The gene Nanog was ranked 11th in the embryonic stem cell microrarray data according to the pluripotent profile shown in Figure 5.3.

Figure 5.6. The 10 top-ranked genes obtained from gene profiling with the ‘Sox2’ profile: Sox2 has a higher gene expression level at day 0 compared with days 6 and 9, equivalent expression for days 6 and 9, with the expression level for day 3 lying in between those of days 0 and 6 to 9. Sox2 is ranked 4th in this set of 10 statistically significant genes.

research framework (see Truths I and II, Section 3): there are open questions everywhere at the new biological research frontier. In general however, the challenge for statistical research is harder. When statisticians are collaborating on nonstandard, complex problems
specific to the local context (and almost all local, major bioinformatics projects are nonstandard) carefull consideration needs to be given to whether it would be more appropriate to make use of existing statistical methodology, or whether the development of new statistical techniques is required. Often, a combination of both approaches is needed, and biostatistics has much to offer in this regard. For instance, there is considerable scope for application of Peter Armitage’s sequential trials methodology in gene expression and bioinformatics studies. For example, in microarray experiments in which limited amounts of mRNA or tissue samples are available, a fully sequential (with small $n$) group-sequential or adaptive design in which a small set of slides is preserved for a follow-up experiment, could prove efficient (see, for example, [38]). There will always be practical limitations on sequential methodology applied in the bioinformatics setting, just as there are in clinical applications, but the potential benefits are yet to be fully explored. The numerous ethical difficulties of combining human experimentation and sequential trials methodology would, for the most part, not arise in the bioinformatics setting.

It is clear that new statistical challenges will continue to emanate from the biology, medicine and public health of the future, from breakthroughs at the interface of statistical genetics, statistical epidemiology and biostatistics, and from areas, events and technologies as yet unknown. Whatever the challenges, guidance from expert statisticians will be needed.

7. Acknowledgements

First and foremost, I would like to thank Peter Armitage for inspiring generations of PhD students, for his contribution to my own PhD and subsequent work, and for his forebearance in allowing me the privilege of including small tokens of his enormous research portfolio in this paper. I have found Peter Armitage’s papers and books breath-taking in their depth and technical clarity. I would like to thank Professor Terry Speed (UC Berkeley and Walter and Eliza Hall Institute, Melbourne) for his unfailing and generous advice in recent years on microarray experiments and bioinformatics issues. His path-breaking work on pre-processing and analysing microarray data, in which he recognised early on that existing (crude) background correction and data pre-processing methods were eliminating much of the biological effect of interest before the data were even analysed, has been a major influence on statistical developments to date. I am also very grateful to Dr Keith Baggerly of the M.D. Anderson Cancer Center (Houston, Texas) for generously providing a critique of the Petricoin et al ovarian cancer study and sequelae and Figures 4.2 and 4.3 of this paper. Finally, I am grateful to Professor Vern Farewell for his editorial critique and contributions to this paper.

References

E-mail address: patty.solomon@adelaide.edu.au