Bioinformatics: A Challenge for Statisticians

Graham J. G. Upton

Department of Mathematical Sciences
University of Essex, Wivenhoe Park, Essex, CO4 3SQ, UK

Abstract

Bioinformatics is a subject that requires the skills of biologists, computer scientists, mathematicians and staisticians. This paper introduces the reader to one small aspect of the subject: the study of microarrays. It describes some of the complexities of the enormous amounts of data that are available and shows how simple statistical techniques can be used to highlight deficiencies in that data.

1 What is Bioinformatics?

In our *Oxford Dictionary of Statistics* (Upton and Cook, 2008) we defined Bioinformatics as "A subject concerned with extracting information from data arising from study of DNA, genomes, etc. The subject interweaves biology, computer science, and statistics." As with any dictionary definition this is a definition that provides an impression rather than detailed information. To get a more detailed view of the subject we need to turn to a specialized book. A useful introduction to the subject is provided by Lesk (2008). His book begins as follows:

> "Biology has traditionally been an observational rather than a deductive science.Although recent developments have not altered this basic orientation, the nature of the data has radically changed. ... in the last generation the data have become ... much more quantitative and precise."

Thus the word "data" appears immediately! Lesk goes on to write that

> "... an obvious property of the data of informatics is their *very very large amount*"

The italics are those of Lesk and they emphasise the key property that makes the subject a challenge for statisticians. There is a parallel with a National Lottery: the chance of anyone winning is very small indeed and yet, in most weeks, there is a winner — seemingly paradoxically, rare events are happening constantly. So it is with Bioinformatics! Anything that can happen, will happen — ordinary notions of significance must be abandoned.

The word 'Bioinformatics' is an amalgam of 'Biology', the science of living things, and 'Informatics', the science of information, information processing and information systems. Informatics leans heavily on computer science and many introductory books on Bioinformatics place it in the context of some form of computer programming (Gibas *et al*, 2001; Orengo *et al*, 2002; Tisdall and Stein, 2003; Gentleman, 2008). In the introduction to their book Gibas *et al* write that

> 'The field of bioinformatics relies heavily on work by experts in statistical methods and pattern recognition.'

Thus the impact of statistics is clearly recognized. Nevertheless, it is not sufficient to understand the statistical methodology — one must also understand the biological implications. Gibas and Jambeck warn that

> "These new tools [algorithms, databases, statistical tools] also give you the opportunity to over-tinterpret data and assign meaning where none really exists."

At Essex we are constantly performing 'apple tests' for situations where we have discovered apparent links between phenomena: we ask the questions 'Are these results repeatable?' and 'Was this certain to happen in any case?'. Generally, and disappointingly, the answers are usually 'No!' and 'Yes!'.

1.1 The growth of Bioinformatics

For academics an obvious measure of growth lies in the number of journals that are devoted to some aspect of the discipline. In the case of Bioinformatics this s a revealing exercise, as Table 1 illustrates.

Clearly something occurred at the start of the second millenium that sparked activity in the subject. It is not difficult to discover that the spark was the completion of the Human Genome Project, which was an international project to chart the entire genetic material of a human being, the first draft of which was completed in 2000. To have some understanding of this project requires familiarity with a number of biological terms.

2 Some relevant biological terms

Most of the following descriptions have been taken from the *Oxford Dictionary of English* (Soanes and Stevenson, 2005): this will be indicated by ODE in the text.

DNA is now well-known as the compound in the human body that contains the essence of an individual. It is a shorthand for deoxyribonucleic acid. The ODE describes it as

Table 1: First publication dates of major Bioinformatics journals

Bioinformatics	1985
BMC Bioinformatics	2000
Briefings in Bioinformatics	2000
Statistical Applications in Genetics and Molecular Biology	2002
Applied Bioinformatics	2004
Journal of Integrative Bioinformatics	2004
IEEE/ACM Transactions on Computational Biology and Bioinformatics	2004
International Journal of Bioinformatics Research and Applications	2005
Current Bioinformatics	2006
International Journal of Data Mining and Bioinformatics	2006
The Open Bioinformatics Journal	2007
Journal of Computational Intelligence in Bioinformatics	2008
Journal of Proteomics and Bioinformatics	2008

> "A self-replicating material which is present in nearly all living organisms as the main constituent of chromosomes. It is the carrier of genetic information.
>
> (Each molecule of DNA consists of two strands coiled round each other to form a double helix, a structure like a spiral ladder. Each rung of the ladder consists of a pair of chemical groups called bases (of which there are four types), which combine in specific pairs so that the sequence on one strand of the double helix is complementary to that on the other: it is the specific sequence of bases which constitutes the genetic information.)"

In DNA the four bases are *Adenine* (A), *Cytosine* (C), *Guanine* (G), and *Thymine* (T). Adenine and Guanine contain three Nitrogen atoms and are called *Purines*, while Cytosine and Thymine have just two Nitrogen atoms and are called *Pyrimidine*. It is the specific sequence of bases that distinguishes individuals from one another (though, for the most part, the sequences are at least very similar).

Alongside their DNA, organisms have RNA, described in the ODE thus:

> "Ribonucleic acid, a nucleic acid present in all living cells. Its principal role is to act as a messenger carrying instructions from DNA for controlling the synthesis of proteins, although in some viruses RNA rather than DNA carries the genetic information."

Adenine, Cytosine and Guanine appear (in the same order that they appear in an individual's DNA) in RNA, but in RNA Thymine is replaced by another Pyrimidine: *Uracil* (U).

Familiar terms to the layman are *chromosomes* and *genes*. The ODE defines a chromosome thus:

> "A thread-like structure of nucleic acids and protein found in the nucleus of most living cells, carrying genetic information in the form of genes.
>
> (Each chromosome consists of a DNA double helix bearing a linear sequence of genes, coiled and recoiled around aggregated proteins (histones). Their number varies from species to species: humans have 22 pairs plus the two sex chromosomes (two X chromosomes in females, one X and one Y in males). During cell division each DNA strand is duplicated, and the chromosomes condense to become visible as distinct pairs of chromatids joined at the centromere.)"

and a gene is described technically thus:

> "A distinct sequence of nucleotides forming part of a chromosome, the order of which determines the order of monomers in a polypeptide or nucleic acid molecule which a cell (or virus) may synthesize."

This definition leads us too far into biology for the purposes of this paper, so we will rely on the alternative idea that (perhaps) 'statistical ability is in my genes'.

We noted that the Human Genome Project had a major impact on Bioinformatics. Its purpose was to determine the typical sequence of bases (C, G, A and T) in the human genome and to identify the sub-sequences (of around 100 million bases) corresponding to chromosomes and, within them, the sub-sub-sequences corresponding to about 25,000 genes.

3 Microarrays

The Bioinformatics work in our department at Essex has concentrated on data from microarrays, and the remainder of this paper will concentrate on these objects.

Microarrays are manufactured by several companies and more than one technology is involved: we will concentrate on the so-called GeneChip microarray produced by Affymetrix, which is the leading manufacturing company. The HGU_Plus2 microarray is the most recent Affymetrix microarray designed to provide information about human tissue. It consists of a lattice of regions with 1164 rows and 1164 columns (the *array*) all contained on a thin wafer the size of a small postage stamp (hence *micro*array).

At each of the 1.4 million locations on the array there are thousands of individually constructed sequences of the bases (C, A, G and T) introduced earlier.

These sequences (created using photolithography) are chosen to match sequences occurring in specific human genes, with the aim being to discover the extent to which that gene is expressed in a given piece of human tissue.

The initial hope was that if an individual human suffered from some unfortunate disease then, by comparing the levels to which genes were expressed for that individual with the level to which they were expressed for a normal individual, it would be possible to identify the (presumably defective) gene responsible, which would open up the possibility of some form of genetic engineering to eliminate the disease. Unfortunately, in reality, when one compares results for two pieces of human tissue, there are inevitably major differences in the expression levels of dozens (if not hundreds) of genes.

In a GeneChip the sequences at each location are 25 bases in length and the collection of sequences at a location is referred to as a *probe*. As previously stated, each probe is intended to match a 25-base sequence in a specified gene, though occasionally the sequence may appear in more than one gene. To guard against the resulting possibility of misinterpreting a high level of expression as indicating the target gene when it may result from some other gene, and to guard against manufacturing flaws or failure to correctly follow the biological protocol for preparation of the microarray, there are typically 11 probes all targeting the same gene (with sequences matching different segments of the gene): these 11 probes are referred to as a *probe set*. The members of a probe set are typically distributed widely across an array, so that if there a small region of the array has flawed data, then most members of a probe set will remain unaffected.

There are several statistically based procedures in common use for accumulating the information from the probes in a probe set — typically testing for, and discarding, outlier values (see e.g. Li and Wong 2001; Wu *et al*, 2004). At Essex we have adopted a rather different approach, attempting to identify incorrect probe values by examining and comparing entire microarrays rather than just the 11 values in a probe set.

4 Comparison of replicate microarrays

A mainstay of experimental design has been the use of repeat observations (replicates) to give an idea of experimental error (uncertainty). Although microarrays are becoming cheaper, they remain an expensive item and technical replicates (i.e. material coming from the same sample) are very rare. Biological replicates (material coming from different samples from a common population) are less rare. The existence of any type of replicate provides an immediate opportunity for a simple test for spatial flaws.

Figure 1 provides an example of how a particularly simple statistical idea can

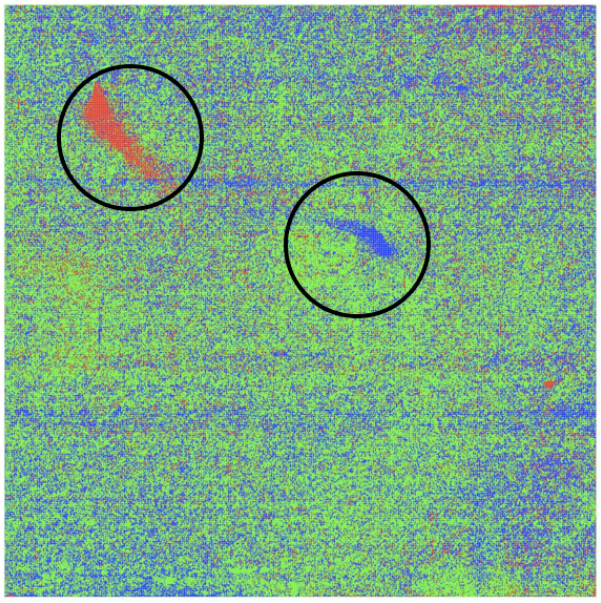

Figure 1: Colour-coded comparison of three replicate microarrays showing undesirable spatial flaws

be very revealing. The diagram presents results from a comparison of three biological replicates. Each of the 1.4 million locations on the microarray has been represented by a dot coloured either red, green, or blue, depending on which replicate provided the greatest value at that location. Whilst there are other features present in the diagram (such as the evident horizontal line 'background' pattern) it is evident that two of the three replicates have significant regions of enhanced (which means incorrect!) signal. A rather more sophisticated approach was given by Upton and Lloyd (2005). A detailed overview of spatial defects in Affymetrix GeneChips is provded by Langdon *et al* (2008).

This simple spatial screening implies that the affected probes can be detected without reference to the values of the other members of their probe sets. However, detection is simply the first step. There then comes the question of what to do next. A simple solution is to treat each as an outlier and to discard it. However, there are only 11 members in a probe set, so discarding nearly 10% of the information would be reckless if there was an alternative. Research at Essex has suggested an algorithm that may be capable of repairing 'damaged' values by reference to the values in the remaining replicates (whilst, nevertheless, maintaining independence of replicate values); details are given by Arteaga-Salas *et al* (2008).

5 Large-scale comparison of microarrays

Biologists carrying out investigations of particular genes, or of particular races or subspecies, naturally focus on the results of their own experiments. However, a rich pool of information is waiting to be tapped: it is called the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository. At the time of writing (early 2009) it contains nearly 300 000 publicly available data sets which include over 20 000 of the HGU133_Plus2 arrays. The size of this database is constantly growing; in late 2007 we had downloaded almost all the arrays then available and that was then fewer than 3000. Nevertheless, 3000 arrays, each containing nearly 1.4 million data points, provides an appreciable data resource!

All the probes in a probe set are supposed to react to the presence of a gene, with the extent of their reaction reflecting the extent to which the gene is present. Across 3000 arrays one can expect a good range of levels of gene expression and thus the usual product-moment correlation coefficient should have a value reasonably close to 1. For the case illustrated in Figure 2 the correlation is 0.76.

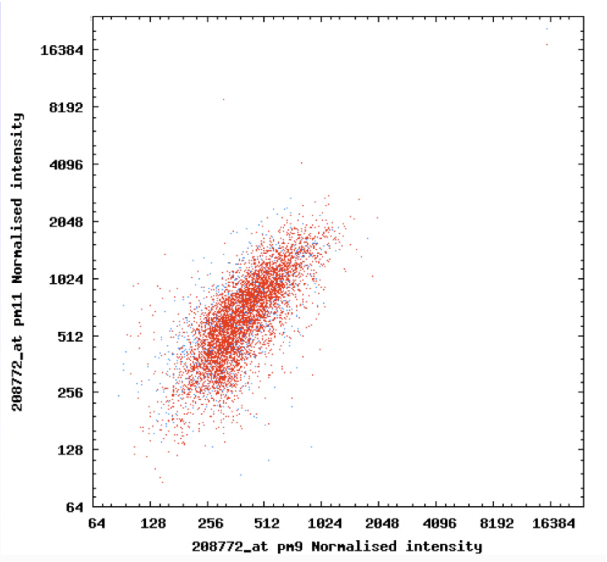

Figure 2: A typical scatter diagram for two correctly functioning members of the same probe set. Note the logarithmic scales (the underlying distributions are approximately log-normal).

Figure 2 presents a clear picture of two correctly functioning probes. However, it is not feasible to study the 55 possible scatter diagrams for the members of a typical probe set. A swifter visual alternative is provided by a *heatmap* of

correlations such as that illustrated in Figure 3. It is immediately apparent that one probe is not behaving in an expected fashion.

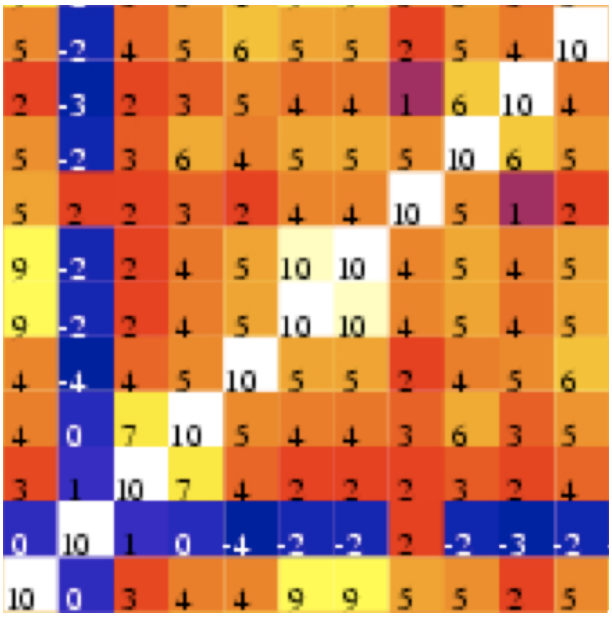

Figure 3: A heatmap showing one member of a probe set that is not well correlated with the other members (which are reasonably well correlated with one another). The numbers are correlations multiplied by 10 and rounded.

Possibilities that occurred to us were that the probe sequence (TCCTGGACTGAGAAAGGGGGTTCCT) corresponded to more than one gene, or that there had been some form of transcription error. We calculated, for all the probes in two related probesets, their correlations with about 245 000 other probes. The result is shown in Figure 4: there are four probes that are conspicuously different to the others (they have correlations that exceed 0.5 with around 15 000 other probes across the microarray). Clearly this is not a genetic effect. The key to discovering the cause was shown by the probe sequences of the four atypical probes: all contained the base sequence GGGG. It turns out that there is a biochemical explanation for why such probes are atypical: the implication is that such probes are not effective and should not be included. For details see Sanchez-Graillet *et al* (2008) and Upton *et al* (2009a).

For the previous situation it was the similarity of the probe sequences that provided the clue to the misbehaving probes. In the next case, there is a very different type of explanation. The problem (and its solution) emerged when we created a composite heat map (Figure 5) for two unrelated genes.

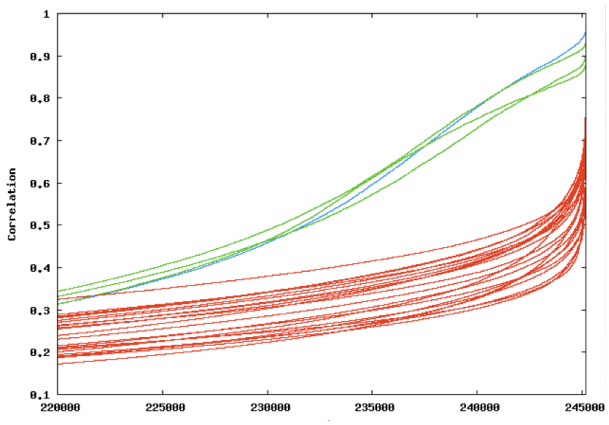

Figure 4: A graph showing, for each of a number of related probes, the right-handend of the cumulative distribution of their correlations with around 245 000 other probes. The top four probes therefore have correlations that exceed 0.5 with nearly 15 000 other probes.

In Figure 5 we see that probe 16 is not well correlated with the other members of its probe set, but it is surprisingly well correlated with probe 1. The answer to the question 'What do these probes have in common?' is that they sit adjacent to the outer edge of the microarray. This outer edge includes an alternating sequence of extremely high-valued probes. Our investigations (described in detail in Upton, 2009b) again used correlation as a discovery tool.

Table 2: The typical average correlation between the values of probe that have large-valued neighbours. The probes are subdivided by the value of their largest neighbour.

ln(Value of largest neighbour)	4.0-	5.0-	6.0-	7.0-	8.0-	9.0+
Median correlation	0.07	0.13	0.25	0.50	0.70	0.85
No. of probes ('000)	893	373	67	14	4	3

Table 2 shows how the typical correlations between the neighbours of large-valued probes is influenced by the magnitude of the large neighbour. Evidently the correlation is greatest for those with the largest neighbours (as with those probes adjacent to the edge of the microarray). To properly understand this phenomenon requires some understanding of how values are assigned to probes.

To quantify the levels at which genes are expressed, a fresh microarray is immersed in a 'soup' containing biological material and a fluorescent marker.

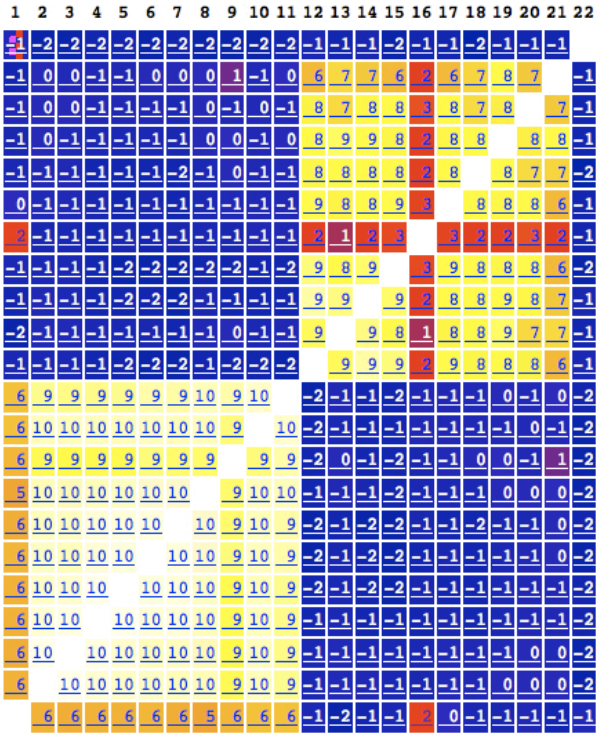

Figure 5: A composite heatmap for two unrelated probe sets. Probes 1 to 11 are well correlated, as are (for the most part) probes 12 to 22. The features of interest concern probe 16, which is poorly correlated with the others in its probe set, but is (relatively) well correlated with probe 1.

Following a set protocol, the microarray is withdrawn and thoroughly washed so that the only biological material that remains is that which has bound itself to the material initially on the microarray. The microarray is then scanned by a laser and the amount of fluorescence at each location provides the quantification that is assigned to the probe and is presumed to indicate the expression level of the corresponding gene.

The laser is therefore effectively 'seeing stars'. When we look at stars in the sky we do not see well-defined circles, but simply regions of light that are brightest in the centre. So it is when the laser looks at the microarray and there are interesting statistical issues (not directly relevant here) concerning the formula used to change the initial laser records into the value reported. What is relevant is that a poorly focused laser will blur all the bright dots to a greater extent than will a well-focused laser. Poor focusing implies that some of the blurred light may be assigned to a neighbouring probe. Since this happens to the same extent

across the entire microarray it follows that these neighbouring values will have the correlated values that are reported in Table 2.

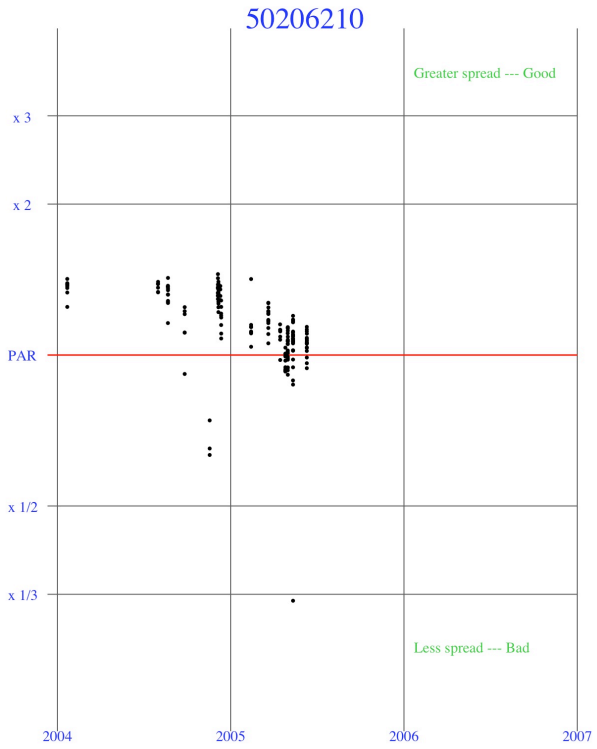

Figure 6: A time series graph for a single laser scanner. The quantity plotted is the simple ratio of the variance of the logarithms of the values on a single array to the median of this variance for all arrays. High values correspond to sharp images and low values to blurred images.

Since each GeneChip data set records both the date at which the data were recorded and the identity of the laser scanner used it is possible to track the performance of scanners over time. Figure 6 shows a previously unreported time series for a particular scanner. Ignoring some obvious outliers it is obcious that this scanner was performing well in 2005, but was starting to deteriorate in 2006. Similar results have been found for other scanners.

6 Summary

This paper will stand out from its peers because of its concentration on context and complete absence of equations. However, the reader should not be misled into thinking that because the techniques reported (the use of the maximum, the median and linear correlation) are simple, that the subject matter is also simple. The difficulties lie in the huge amount of data (that needs to be efficiently manipulated) and the wide variety of sources of variation, some obvious and intended (genetic make-up), some biochemical (the problems with probes containing the GGGG sequences — we believe that there are others such), some due to the construction of the microarrays (we have not previously mentioned that each probe is one of a pair, with the second probe intended as a measure of background but often greater than the primary probe) and some physical (the inability of laser scanners to perfectly measure the light intensity of microscopically small light sources).

Microarrays are growing ever more varied and their capabilities are being extended. The latest arrays from Affymetrix use smaller sized features to provide information on more than six million probes simultaneously. Accurate measurement is therefore more challenging and the need for statistical techniques is enhanced.

This paper has concentrated on microarrays because that is the region that the Essex team has been addressing. However, it should be realised that any book on Bioinformatics will be addressing only a few pages on this branch of the subject. There is plenty of scope for mathematicians of all types (both pure and applied) in addressing the many challenges that Nature has presented.

Acknowledgements

Just as Bioinformatics is a discipline that fuses ideas from different areas of science, so effective research requires expertise in different areas. The work reported here has benefitted hugely from the computational expertise of my former colleague Dr. Langdon, the biological expertise of my colleague Dr. Harrison, the industry of my student Mr. Arteaga-Salas, and the efforts of the others in the Essex bioinformatics team.

References

Arteaga-Salas, J. M., Harrison, A. P., and Upton, G. J. G. (2008) Reducing spatial flaws in replicate oligonucleotide arrays by using neighbourhood information. *Statistical Applications in Genetics and Molecular Biology*, **7**, 29.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C.,

Kim, I. F., Soboleva, A., Tomashevsky, M., Edgar, R. (2007) NCBI GEO: mining tens of millions of expression profiles-database and tools update. *Nucleic Acids Research*, (**35** Database), D760-D765.
Gentleman, R. (2008) *R Programming for Bioinformatics* Chapman and Hall.
Gibas, C., Jambeck, P., and LeJeune, L. (Ed) (2001) *Developing Bioinformatics Computer Skills: An Introduction to Software tools for Biological Applications* O'Reilly Media, Inc.
Langdon, W. B., Upton, G. J. G., da Silva Camargo, R. and Harrison, A. P. (2008) A survey of spatial defects in Homo Sapiens Affymetrix GeneChips. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, http://doi.ieeecomputersociety.org/10.1109/TCBB.2008.108.
Lesk, A. M. (2008) *Introduction to Bioinformatics*, 3rd edn. OUP Press.
Li, C., and Wong, W. H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, **98**, 31-36.
Orengo, C., Jones, D., and Thornton, J. (Eds) (2002) *Bioinformatics: Genes, Proteins and Computers*. Taylor and Francis.
Sanchez-Graillet, O., Rowsell, J., Langdon, W. B., Stalteri, M., Arteaga-Salas, J. M., Upton, G. J. G., and Harrison, A. P. (2008) Widespread existence of uncorrelated probe intensities from within the same probeset on Affymetrix GeneChips. *Journal of Integrative Bioinformatics*, **5** (2):98.
Soanes, C. and Stevenson, A. (Eds) (2005) *The Oxford Dictionary of English*, (revised edition). OUP Press.
Tisdall, J. D., and Stein, L. (2003) *Mastering Perl for Bioinformatics* O'Reilly Media, Inc.
Upton, G. J. G. and Cook, I. T. (2008) *Oxford Dictionary of Statistics*, 2nd Ed. revised. OUP Press.
Upton, G. J. G., Langdon, W. B. and Harrison, A. P. (2009a) GGGG sequences cause incorrect measurement of gene expression in short-oligo microarrays. *BMC Bioinformatics*, (to appear).
Upton, G. J. G., Andrade-Pacheco, R., Pérez-Torres, L., Langdon, W. B., and Harrison, A. P. (2009b) The effect of blur on probe values in Affymetrix GeneChips. *Bioinformatics*, under review.
Upton, G. J. G. and Lloyd, J. C. (2005) Oligonucleotide arrays: Information from replication and spatial structure, *Bioinformatics*, **21**, 4162-4168.
Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., Spencer, F. (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, **99** (468), 909-917.