NPL Report CMSC 23/03

Report to the National Measurement System Directorate, Department of Trade and Industry

New Directions – Software Issues in Bioinformatics

Simon Cowen & Steve Ellison, LGC

May 2003

New Directions – Software Issues in Bioinformatics: recommendations for the Software Support for Metrology programme 2004-2007

Simon Cowen and Steve Ellison
Laboratory of the Government Chemist

May 2003

ABSTRACT

This report sets out the conclusions of a study into measurement issues in bioinformatics that will need to be addressed to meet future challenges within the National Measurement System (NMS). This study is one of five within the "New Directions" project, in the current NMS Software Support for Metrology (SS*f*M) programme. The study has investigated the current status of relevant work in bioinformatics and measurements for biotechnology. The main aim of the study was to identify whether there are measurement issues in bioinformatics which should be addressed by the SS*f*M programme. The report provides input to the process of formulating the 2004-2007 SS*f*M programme.

ISSN 1471-0005

National Physical Laboratory
Queens Road, Teddington, Middlesex, TW11 0LW

Approved on behalf of the Managing Director, NPL
by Dr Dave Rayner,
Head of the Centre for Mathematics and Scientific Computing

CONTENTS

Executive Summary

Introduction

This report sets out the conclusions of a study into measurement issues in bioinformatics that will need to be addressed to meet future challenges within the National Measurement System (NMS). It forms part of the "New Directions" project, in the current Software Support for Metrology (SS*f*M) programme. The "New Directions" project encompasses a wide range of topics: legal metrology, digital signal processing, future directions in mathematics and computing, soft metrology, and the subject of the current report, bioinformatics and measurements for biotechnology. Separate reports are being produced for each of the five topic areas and it is the intention that each report should stand alone and that overlap between the topic areas should be avoided. The intention is that the reports should provide useful input to the process of formulating the 2004-2007 SS*f*M programme. Further information about the SS*f*M programme can be found at the website address: www.npl.co.uk/ssfm/.

The main aim of the study reported here was to identify whether there are measurement issues in bioinformatics which should be addressed by the SS*f*M programme. The field of study covers:

- Applications of data acquisition and processing systems for high volume measurement and test systems.
- Potential problems arising from cross-platform incompatibilities in measurement data format in Bioinformatics.
- The impact of measurement and measurement uncertainties on bioinformatics, including the impact on pattern matching and identification.

The study involved direct contacts with industry, academic and measurement institute representatives, supplemented by an extensive literature study aimed at identifying key areas of technological development and software application in bioinformatics.

Issues in Bioinformatics

As a field, bioinformatics is highly diverse, gaining greater importance in biology, and developing with great rapidity. As users become more familiar with the fusion of computational techniques with experimental procedure, we can expect to see more papers which describe the use of bioinformatics tools in a highly routine manner.

Uncertainty is present and widely acknowledged in bioinformatics data. Some is due to measurement uncertainties; for example, uncertainties in biological reference materials may be large compared to those in physical measurements, and run to run variability in measurements is often large at low levels (as it is in chemical measurement). However, the largest sources of uncertainty in the data do not relate directly to measurements, but to such influences as the inherent variability of biological test materials, their provenance, and the complexity of the biological systems from which responses arise. Thus the overall variability of data means that measurement uncertainty is not a major issue at present in bioinformatics as characterised within this report. However, the reliability of the underlying data will be more influential as the sophistication of these tools increases and data mining extracts more information from data. The issue will also become increasingly important where bioinformatics techniques (high volume sample throughput and data handling) are applied in a regulatory context. Thus, uncertainty issues are likely to increase in importance.

In considering software issues, four different classes of software were identified:

- Database technology,
- Modelling and prediction software,
- Instrument control and
- Statistical software.

Though the state of the art varies from R&D activity to well-developed commercial activity with excellent QA, consistent themes in all these cases are as follows:

- Provision of reliable test data, good design practices which encourage application of design and test disciplines.
- A need for awareness among a rapidly developing programmer and (sometimes) end-user community.

These are recognised issues, and many activities are under way in the bioinformatics and instrument development community. Yet it is not clear how good UK company access is in these fields, and the preponderance of start-ups and SMEs makes it likely that improving access and awareness of existing tools will assist the UK community.

The study found that terminology, vocabulary and data format issues are very substantial in bioinformatics. These issues are, however, well recognised by the bioinformatics community and considerable effort is already under way to address them.

Relevance of SSfM

As we have stated, the existing principles of SS*f*M in its current form relate well to bioinformatics software, and the bioinformatics community (academics, commercial software developers and end users) should be made aware of and encouraged to participate in SS*f*M activity. This is advantageous in that those issues identified as being particularly important in biology (data compatibility, terminology, and integration of diverse data types) would then be included, giving more complete coverage of software and data issues in science as a whole. Specific SS*f*M activities that are relevant include:

- *Data fusion*: Issues in data fusion in other fields should map well onto bioinformatics data interpretation issues
- *Reference data set generation*: SS*f*M's methods for producing reliable data sets for statistical software could provide a useful basis for some type of bioinformatics software validation.
- *Instrument software development and validation*: Existing SS*f*M guidelines and practices could be applied in biological measurement instrument development and application.

Other NMS programmes

The two principal programmes which might benefit from SS*f*M support or collaboration are the Valid Analytical Measurement (DNA measurement) and Measurements for Biotechnology (M*f*B) programmes. The present (2002-2004) M*f*B programme, in particular, is specifically addressing array technology and measurement uncertainties in biological measurement. However, there are few areas of SS*f*M interest which are unique to biological measurement, and numerical accuracy and even measurement software reliability are rarely critical provided that obvious pitfalls are identified and eliminated. The most urgent needs are accordingly to bring

biological measurement practices in general up to those in chemical and other measurement fields.

In the longer term, it is possible that biological measurements will begin to develop approaches and protocols related to uncertainty estimation following ISO guidelines. Where this is the case, it will clearly be important to address the often asymmetric and discrete distributions encountered. These issues are, however, already being addressed within SS*f*M, and provided that they continue to receive support, will be applicable when the need arises.

Recommendations

The key themes arising in this report suggest that SS*f*M activities are already well applicable to biological measurement. However, it is important to be aware that a wider range of statistical procedures is in use than in many measurement areas, that awareness of uncertainty issues is generally far lower, and that measurement instrument control software tends to be both complex and rapidly developed.

These conclusions suggest the following general recommendations in developing SS*f*M:

- Assess and, if necessary, act to improve the quality and time-to-market of biological measurement software provided by UK companies. This should involve a sector-specific review of the current state of the art among the UKs bioinformatics measurement software developers which reviews the software specification and development process; algorithm use and implementation; appropriateness of algorithms; numerical accuracy; and generation and use of test data. The review should make recommendations for provision of guidance, software tools and training.
- Make provision for communication on measurement uncertainty issues at least with those UK measurement institutes involved in measurement standards provision and preferably with bioinformatics community representatives, to identify specific needs for guidance and to address those needs in collaboration with other relevant NMS programmes.
- Publicise SS*f*M's activities on reference data sets with a view to a) stimulating the adoption of similar principles for reliable bioinformatics test set development, and b) in the longer term, identifying and supporting test set development projects. For bioinformatics, this can be best achieved through development and application of pilot test sets for software intended for statistical analysis of large data sets, including sets with discrepant values.
- Review statistical tools (such as data mining tools) emerging from bioinformatics research to assess their applicability to other measurement fields and to metrological development, and conduct feasibility studies on prospective applications.
- It is recommended that industrial contacts focus in the first instance on DNA microarray technology, as this is the best developed field and a good model for future interaction.
- This report and its recommendations should be made available to key UK organisations, particularly NIBSC, to stimulate additional input to the SS*f*M formulation process.

New Directions – Software Issues in Bioinformatics

1. Introduction

1.1 Background

This report is intended to inform future Software for Metrology and related projects on software relevant to the NMS programmes. Biological measurement is among the new fields of measurement introduced into these programmes, and it has become clear that software, in the form of bioinformatics, plays a key role in developing and extending knowledge of biological systems. It is accordingly necessary to examine the role of bioinformatics in biological measurement, and in particular to identify any issues arising either from the nature of the software used, or from measurement issues relevant to its application.

Laboratory science has for many years had mechanisms in place which enable the quality and accuracy of experimental and analytical data to be assured, and the Valid Analytical Measurement (VAM) programme [50] is the principal scheme in the UK. No programme has, however, yet been put in place to address similar measurement and accuracy issues in bioinformatics and bioinformatics software, and as the uptake of and reliance on bioinformatics and large-scale data processing grows, such a study is becoming more important.

1.2 Scope of the study

Bioinformatics has been defined [1] as "the mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information", in other words, primarily the analysis and data mining of genomic and proteomic information. However, this is a somewhat narrow definition. The US National Institutes of Health recently proposed the more inclusive "research, development or application of computational tools and approaches for expanding the use of biological, medical, behavioural or health data, including those to acquire, store, organise, archive, analyse or visualise such data" [2]. Currently, to judge from the literature, most emphasis is placed on the extraction of the maximum value from a given data set, more than the generation of new data itself.

Specifically excluded are those techniques of computational biology such as the modelling of molecular structure, dynamics and interaction. Using theoretical methods, mathematical modelling and simulation, they fall outside the accepted definitions of bioinformatics as outlined.

The range of applications of bioinformatics is therefore substantially wider than the remit of the NMS software or measurement programmes. The scope of the present study is accordingly limited to the identification of issues associated with measurements for bioinformatics, rather than with software issues in bioinformatics as a whole. In particular, the study is intended to cover the following:

- Applications of data acquisition and processing systems for high volume measurement and test systems.
- Potential problems arising from cross-platform incompatibilities in measurement data format.

- The impact of measurement and measurement uncertainties on bioinformatics, including the impact on pattern matching and identification.

This project, funded under the SS*f*M programme, represents the first investigation into the measurement software issues affecting the end user of bioinformatics. The report accordingly provides a general description of bioinformatics as a whole, followed by a discussion of particular applications of software, the types of software in use and the software issues arising. It concludes with a consideration of the issues arising in measurements associated with bioinformatics.

2. Conduct of the study

2.1 Consultation

Approaches for consultation were limited to individuals or organisations with existing contact with LGC or NPL who had previously agreed to discuss biological measurement issues. A list of organisations approached is included as Appendix 1.

A focus group was felt to be an effective way to gain an understanding of current thinking in the topic, and the best approach for this study. Accordingly, a meeting was organised and held at LGC in October 2002, based around presentations on the major issues, followed by facilitated discussion. The participants included interested parties from academia and industry and, being organised on the same day as another forum on biological measurement uncertainty, relevant input was also gained from the latter subject material. A list of those attending is given in the appendix. The discussion from this focus group is summarised in Appendix 2.

2.2 Literature search

A review of the general literature in bioinformatics over the last three years was also undertaken, in order to try to extract the issues most relevant to measurement, and to identify particularly strong themes prior to making a recommendation on future directions and project funding.

The literature in bioinformatics has grown steadily over the last ten years, and now accounts for some 2% of all the articles contained in the PubMed literature database [3]. This coincides with its maturation into a field of study in its own right, as well as the expansion in the genome sciences. The increasing use of information technology in biology, and the corresponding greater need for sophisticated data analysis and mining techniques also drive research in this area, as has already been stated. Thus, there is a wealth of bioinformatics literature describing both the development and application of tools, generally falling into one or more of the following categories:

- Genome sequence determination and analysis
- Gene prediction algorithms
- Microarray data analysis
- Computational and homology modelling of protein structures
- Comparative genomics and biochemical pathway deduction
- Structural and functional genomics
- Evolutionary analysis
- Gene function in disease
- Other techniques (mass spectrometry, microscopy) with large data sets

The biology literature is also beginning to report on work carried out by the end user; one who is not a developer, but uses these tools to perform tasks such as comparing query nucleotide and protein sequences with the contents of central databases, the aim being to aid in drawing conclusions of biological significance from the data obtained. This is the end point and 'routine'

aspect of bioinformatics, which is now experiencing more general uptake. This latter literature is not so extensive, and the question of the relation between measurement accuracy and bioinformatics has not received attention to date.

2.3 Individual contacts

Meetings were also held with individuals and groups practising in bioinformatics, in order to gain a more detailed set of perceptions, views and requirements. A report on one such meeting, with the statistics department at Unilever Research, is attached as Appendix 3.

We note here that in practice, individual contact has proven more difficult to arrange and follow through than envisaged, and although a substantial list of voluntary contacts was invited to comment, relatively few responded either with comment or in person. Greater weight has accordingly been given to the outcome of literature reviews.

Given the relatively low level of response in this study, it is clearly important that further efforts are made during SS*f*M formulation to consult with organisations who have not contributed directly to the present study. Appendix 1 lists organisations approached for comment. While most key bioinformatics organisations in the UK have responded, a notable absence to date is NIBSC, an organisation with key responsibilities in standardisation for biological measurement and with a memorandum of understanding with NPL for National Standards maintenance. It would be particularly valuable to obtain their views on the measurement and measurement uncertainty issues raised in this context.

3. General description of bioinformatics

As the power of information technology has continually increased, together with a significant real-terms drop in costs, it has been applied to more areas of science. Although historically important in the physical sciences and engineering, computing has only relatively recently been extensively used in biology, and there is still much potential for its use. There have been two major effects on the study of biological systems. Firstly, there is now a much greater capability to generate, store and process large amounts of data, and this has produced a number of issues around how to validate and interpret such data. Secondly, increased computing power has given biologists the means to create and run complex software designed to extract well-hidden information, or to make statistically valid predictions of the properties of a system. Both are now recognised as major requirements and drivers for the future in the life sciences, and they require scientists with skills in both biology and computing.

This new skill set is generally referred to as 'bioinformatics', and although the term was first used in the 1960s by Rybak [4], bioinformatics only emerged as a mature discipline later, when the first large-scale gene sequencing programmes began. Since then, the subject has expanded to include the whole interface between biology and information technology, although there is some debate about what bioinformatics specifically includes. Figure 1 shows the main elements which are currently considered to be core to the subject.

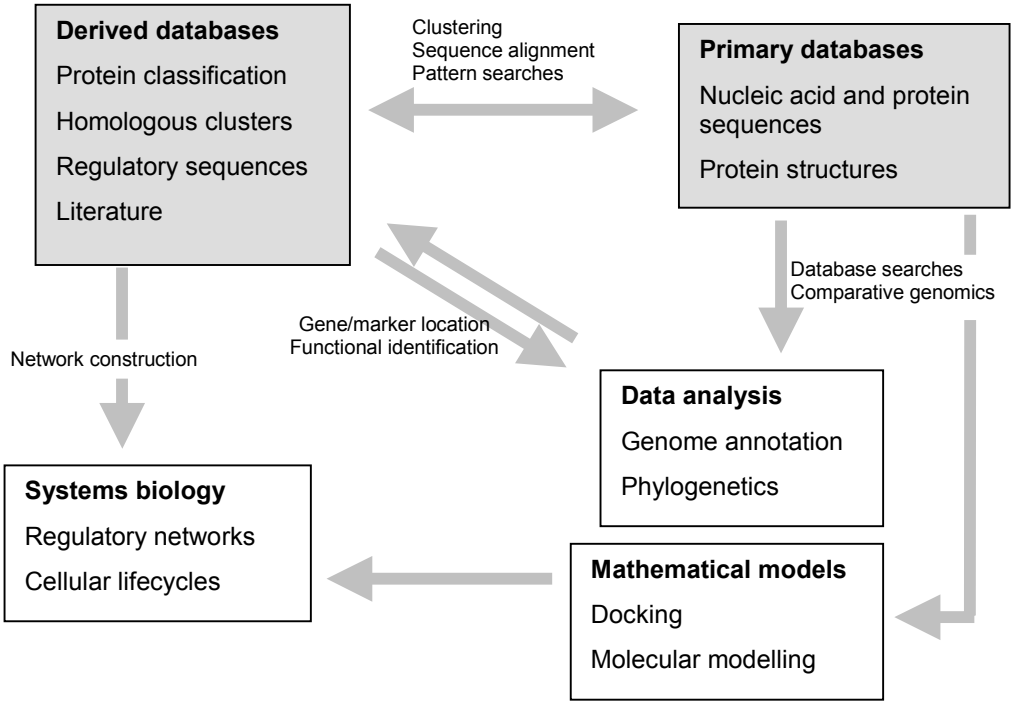


Figure 1: A view of computational biology and how the principal aspects relate to each other

3.1 Techniques in bioinformatics

Without question, the most important application to date of bioinformatics as outlined above is in the characterisation and study of nucleic acids and proteins. The development of powerful sequencing technology and the use of techniques for determining the crystal structure of

biological molecules have required the use of computing power to an unprecedented degree over the last three decades. For example, with the establishment of the Protein Data Bank in 1971 [5] and GenBank in 1982 [6], the ability to search and query structures and sequences became available to the biologist, who could then uncover homologies and analogies, and deduce evolutionary and functional relationships. As the amount of data has grown, this approach has become more powerful, and new areas of research have appeared, including:

Whole genome sequencing - the first genomes to be sequenced, those of viruses, required only a few thousand bases to be determined, but larger organisms, particularly those of mammals, are several billion bases long. Seen as particularly valuable for medical and pharmaceutical research, the human [7,8] and mouse [9] genome sequencing projects have required significant computing power to piece together the entire sequence from the many smaller fragments produced in the laboratory.

Structural and functional genomics - the construction of detailed genetic and physical maps of the genome, and the use of this information to determine experimentally the relationship between gene characteristics and protein function are the goals of the sequencing projects. It is here that the requirement for high data throughput and sophisticated statistical analysis techniques are required.

Parallel gene expression profiling - new technologies such as microarrays allow the entire mRNA content of cells to be compared across different tissues and under different conditions. It is now feasible to imagine an experiment in which the expression levels of the c.20-30,000 genes in the human genome are monitored at the same time.

Prediction of protein structure from sequence information - the number of known protein structures is significantly lower than the number of genes which have been sequenced. Since the real value of genetic information is in the proteins for which it codes, determining as many protein structures as possible is a priority. One of the most profitable current methods is the use of predictive algorithms which calculate likely forms based on the propensity of particular secondary structures to contain particular patterns of amino acids. However, the results are much improved if protein databases are searched for homologous proteins which, as well as preserving some sequence identity, tend to preserve the overall fold. Therefore, database queries and complex sequence alignment techniques are increasingly necessary.

All these areas are growing in scope and in complexity with the amount of stored information available, and the maturation of the associated information technology and analysis techniques has had to keep pace with the expansion of the science. Bioinformatics is thus an evolving discipline with many different aspects to consider.

However, while bioinformatics is a means of handling, assessing and analysing information, there are circumstances in which it impacts on the actual measurement (of the underlying information), and this will need to be taken into account when considering the validity or accuracy of that information. Examples where this can occur include:

- Image and pattern processing algorithms
- Data storage formats and cross-platform connection and conversion
- Data normalisation and standardisation
- Data mining and cluster analysis techniques

The main techniques used in bioinformatics and their applications are shown in Table 1.

Technique	Purpose
Database searching: Sequence alignment Gene ontology	 Identification of homologues Related documents, distant homologies
Statistical methods: Analysis of variance Significance testing Bayesian statistics Hidden Markov models	 Separation of error sources Sequence analysis
Principal component analysis of multivariate data	Determination of most significant variables in a data set
Clustering: Nearest-neighbour Agglomerative	 Grouping genes with similar behaviour
Dynamic programming	Solution of large-scale numerical problems (e.g. alignment of sequences, used in database searches)
Neural networks	Determination of networks of interactions
Decision trees	

Table 1: Bioinformatics techniques

3.2 Driving factors

In academic terms then, bioinformatics is essentially the development of a set of tools which enable underlying behaviours and relationships between biological systems which were previously invisible to be explained and predicted, but there are significant industrial (commercial) drivers too. In particular, the huge scale on which certain industries operate presents some unique challenges, and in the case of the pharmaceutical industry (the best example to date), the large-scale screening programmes which have to be carried out to develop new drugs illustrate this need well.

The modern drug discovery process relies on large compound libraries which are screened for activity against a specific target based on a disease phenotype or pathology, using high-throughput techniques. Libraries are large but focused, their member compounds being synthesised with systematic variation in structure. Although this combinatorial approach is much more efficient than older, compound design-and-test techniques, the cost of developing a new drug still stands at some $500 million over a cycle time of 10-12 years [10] and continues to increase. Consequently, the techniques and instrumentation used in the industry are operating close to their limits of performance, and alternative ways of identifying compounds of interest will need to be found [11] if pharmaceutical companies are to remain competitive. One such approach is pharmacogenomics, which has been made possible with the large-scale genome sequencing projects undertaken over the last few years. The aim of pharmacogenomics is to understand the relationship between an individual's genotype (inherited genetic characteristics) and the body's reaction to and metabolism of pharmaceutical compounds. Through more targeted discovery and pre-screening of subjects, adverse drug reactions and the time required for clinical trials should be reduced significantly. All this depends on the ability to mine human genome sequence data for polymorphisms and other markers and to correlate clinical studies with genotype in a detailed manner, and very large data handling capabilities are needed.

Other industries also make widespread use of bioinformatics. For example, the consumer products industry has also appreciated the potential that genomics offers in terms of product tailoring and is already conducting gene expression studies on the ageing effects of the sun on skin, for example [12]. Health and personal care, 'lifestyle management' and specialist foods all have different effects on different individuals, and the ability to link this to genotype should help to maximise efficacy and eliminate adverse reactions. As bioinformatics techniques become more established, this will be an increasing trend in many areas.

3.3 Commercial developments and outlook

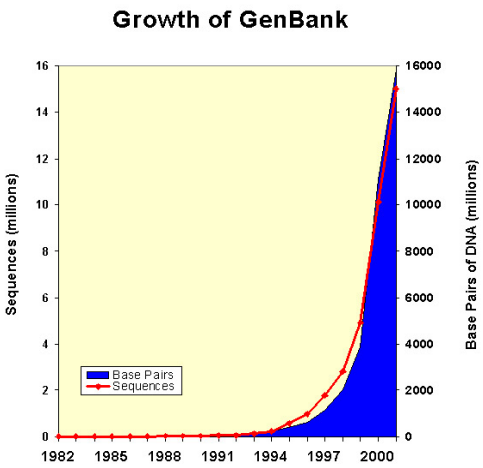
Bioinformatics is closely tied with biotechnology and as such, has suffered a drop in investment and company performance as the high-tech boom has come to an end. This has been seen in reductions in recruiting by large end users, an increase in start-up business failures and a tendency to rethink business models in terms of bioinformatics as a standalone service. However, rather than becoming purely a service to be provided by specialist companies, a significant part of bioinformatics is rapidly being absorbed into biological science in general. For example, more instrumentation is being supplied which is designed to collect and analyse large amounts of data, and the techniques of bioinformatics are included as part of the operating software and procedures. Microarray scanners provide the best current example of this trend, although mass spectrometry is also heading in the same direction, especially with regard to proteomics.

On the other hand, large data sets which have been mined and curated are a saleable commodity, and a business model based on proprietary information, whether from self-generated or a mixture of privately and publicly held data, has arisen - particularly connected with genome sequences. For example Celera Genomics [13], the company set up by the PE Corporation to sequence the human genome for commercial exploitation in direct competition with the publicly funded programme, sells a database containing many genetic markers, expressed sequences and genes not in the public domain version of the genome. This competition between public and private effort is common across bioinformatics in general, and extends across both data and tools. For every commercial product one can conceive, there will be at least one free (at least for academic use) alternative which may well use more cutting-edge technique. Success, therefore, depends on offering a *validated* product which provides trouble-free operation and extensive support.

Nevertheless, the need for skilled bioinformaticists remains, and although the economic downturn and a settling down of the industrial requirement to recruit in this area have had an effect on job opportunities, there are still many openings in the academic sector and in industry for those with 'hard' laboratory skills as well. This last point is important, because with more integration of laboratory and *in silico* science, life scientists must have a greater appreciation of the mathematical and computing aspects of experimental design than ever. It must also be remembered that much of the deductive and predictive nature of bioinformatics work does not give a final answer, but must be verified (and validated where necessary) in the laboratory.

4. Trends in bioinformatics

4.1 Data quantity



Source: GenBank

Figure 2: Growth of the GenBank nucleotide sequence database

The quantity of biological data stored in central databases has increased enormously over the last five years (Figure 2). This in itself has brought new challenges for storage and interpretation, but the fact that the nature of biological data presents specific issues is also a factor to take into account. For example, Schweigert *et al* [14] have described the difficulties present in biological data of ensuring semantic integrity in a database, and the resulting complexity of the problem. Issues around data storage and format are covered in section 5.3.

The second relevant issue is that of the raw data itself and how it is processed. Modern biological measurement techniques frequently involve the acquisition of large numbers of individual units of information (data points or sequence units), and drawing useful conclusions is a matter of robust statistical analysis and replication where appropriate. The measurement workshop held in October 2002 also highlighted the importance of complexity (a large number of variables) and connectivity (interdependence between different data types) in biological data, and these two issues are of fundamental significance in bioinformatics. This, combined with the treatment of raw data, presents a significant validation problem which, in some emerging new fields, will need to be urgently addressed in the near future.

4.2 Application trends

As a recognised, named field of study, bioinformatics as introduced in section 2 is now some ten to fifteen years old, and has undergone rapid change within the last few years. Emerging from the already established field of medical informatics [15], bioinformatics originated mainly as a set of computational techniques applied to the study of the large data sets being generated by the human and other genome sequencing programmes. Hence today, there is still a general link placed between bioinformatics and the genome sciences, and most applications are in fields such as sequence and genome analysis, gene expression profiling and comparison, and (more recently) proteomics.

This rapid change is probably the reason why it is difficult to define bioinformatics in a simple way: it is a discipline which constantly shifts its emphasis from one subject to another. For example, in addition to the core topics of sequence and genome analysis, new experimental methods are emerging in which bioinformatics plays a dominant role, while other, more 'traditional' areas are moving out of focus. For example, new and renewed areas of interest include:

- Modelling and prediction of enzyme kinetics and cellular processes - the genome sequence and proteome provide only a template for the construction of components, and the real key to how living systems work requires an understanding of the many metabolic pathways and kinetic processes in the cell. Known as 'systems biology', this field is increasingly using mathematical and computational techniques to model cellular functions in detail.
- Gene expression and protein arrays - the study of parallel gene expression and the protein products allows the effect of different conditions on thousands of genes to be measured simultaneously. The resulting profiles can then be clustered together and related to phenotype and physical expression. This type of research has been given the term 'phenomics'.

Others, which have hitherto been the subject of a good proportion of bioinformatics research, include:

- Sequence assembly - whole genome sequencing methods require genomic DNA to be fragmented and the individual fragments sequenced separately. The genome sequence is then reconstructed using the information contained in overlapping regions. Although a complex task, several tried and tested methods such as PHRED/PHRAP [16] are now in place. An alternative approach has recently been reported however [17], which appears to offer an improvement over the sequence - overlap - consensus method adopted to date.
- Prediction of protein function from structure - as the number of known protein structures and sequences held in databases continues to grow, homology modelling is expected to be the most reliable method of assigning likely functions and folds to many of the newly identified proteins.
- Molecular modelling - now considered by many to lie in the separate field of computational biology (which is itself a superset of bioinformatics).

These applications are examples of topics which are either essentially 'solved', or which have been overshadowed by more recent approaches. As information technology and the main structure and sequence databases evolve, however, particular areas are likely to move in and out of focus.

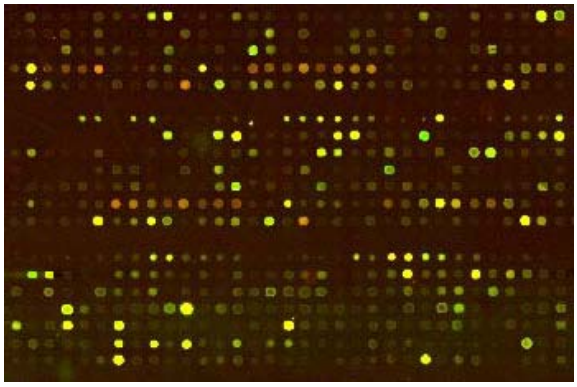
4.3 Applications

Bioinformatics is in some ways far removed from the process of obtaining experimental data in the laboratory, being primarily concerned with the identification and elucidation of patterns in and correlations between the data. However, two factors should be borne in mind. Firstly (and clearly), the quality and usefulness of any conclusions drawn from any computational analysis is directly related to the quality of the underlying data. Secondly, bioinformatics is in most cases able to give only a prediction or deduction with an accompanying statistical measure of reliability, not a definitive conclusion. Such results must be tested experimentally, and in order to ensure that laboratory resources are targeted in the most efficient way, again the initial data must be of the best quality obtainable. Acquiring and presenting this data is likely to involve processing raw data through controlling software, and the validity of the software itself is also included in the need to ensure best practice. This data quality issue has a particular impact on several important bioinformatics-related research themes and in at least one case, for microarrays, has already been raised in the literature [18]. Bioinformatics deals with the storage and representation of data, and there are a number of areas which are not related to

measurement directly, but where measurement is relevant in an indirect way. These are concerned with two important activities: the use of the stored information to make predictions about the biological systems it represents, and the nature of such databases, which is largely defined by the complexity of biological data itself. This section reviews the important aspects of databases and bioinformatics software which use sequence and expression data obtained in the laboratory.

4.3.1 Microarrays

Microarray ('biochip') chip technology is an innovation which has begun to revolutionise the study of gene expression in living systems [19]. With its massively parallel approach, thousands of gene sequences can be contained on one chip substrate and the sample exposed to all simultaneously. The basic underlying principle is as follows: libraries of gene sequences from the organism(s) of interest are created, either as cDNA clones spotted onto the surface of a glass slide, or, in an alternative embodiment, synthesised on the substrate as oligonucleotides using a lithographic process [20]. The selection of genes is tailored to the nature of the experiment. In order to perform an assay, mRNA (which is produced from expressed genes through transcription) is isolated from the sample tissue or cell preparation, and reverse transcribed to cDNA. The cDNA is then labelled with a fluorescent dye. A second cDNA sample is prepared from a reference mRNA source suitable for use as a control (for example a normal tissue sample for comparison with cancer sample) and labelled with a dye of different wavelength.



Source: CSIRO Mathematical and Information Sciences, at www.csiro.au

Figure 3: Microarray fluorescence pattern. The dyes are green (Cy3) and red (Cy5). Yellow spots show genes expressed in both sample and control

The two labelled samples are then mixed and applied to the microarray, where they competitively hybridise with (bind to) the cDNA on the chip. The result is a pattern of spots of different colours, depending on whether one or both of the samples are present (Figure 3).

Quantitative measurements are obtained by measuring the ratios of the two signals at each spot, this figure (which is normally expressed as a base 2 logarithm) providing a measure of the differential gene expression at that point.

Considerable data processing is then required to obtain a set of intensity ratios, which give a measure of the mRNA level present (and hence gene expression level). Indeed, this stage is arguably one of the most important and pressing issues of bioinformatics at the present time. The principal issues surrounding this type of measurement are as follows:

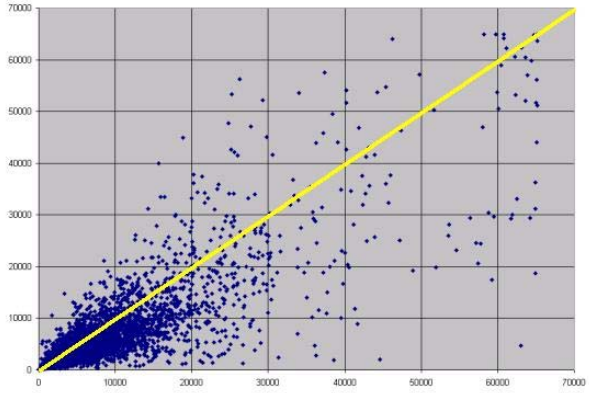
- Processing and normalisation of raw data
- Production of suitable confidence limits on expression levels
- Derivation of expression profiles

- Comparison of inter-laboratory array data
- Data storage formats

Many different methods and software packages are available to carry out these tasks, although the last is really a matter of standardisation (section 5.3). The two stages in generating reliable expression data are described in the next sections.

4.3.1.1 Image analysis

The first stage is to convert the spots on a raw image into an experiment/control intensity ratio for each gene. The standard image format is 16-bit TIFF. Gridding and segmentation are first applied to assign spot coordinates and to classify pixels as spot or background respectively. Techniques such as circles of fixed [21] or variable [22] size or variable shapes [23] are used to overlay a grid onto the image and separate signal from background. The intensity is then usually calculated by subtracting a background value from the mean or median pixel intensity; this may be global or (more commonly) local, and the mean or median can be used. The most appropriate method has not been definitively established, and results seem to vary strongly with the method used [24]. These methods work well when the spot intensity is not saturated, and when the background intensity is less than that of the spot, but there are also methods of dealing with saturated, or other high-intensity spots, based on more complex statistical approaches [25].



Source: University of Texas Southwestern Medical Center, at pga.swmed.edu

Figure 4: Background-subtracted gene expression data taken using a microarray

Complications such as 'comet tailing' and 'doughnut' spotting also occur quite frequently, and a suitable system of quality control is required to reject or accept spots as appropriate. These are usually implemented through cut-off measures and spot shape analysis.

The type of plot produced from an image is illustrated in Figure 4, in this case with arbitrary units rather than explicitly stated logarithmic intensities. As expected, there is a high degree of correlation between the two channels, as most genes do not change their expression levels, but the intensity levels are clustered at the lower end, and scatter is high. The basic problems associated with microarray data are ensuring the comparability of different hybridisations, since each array is used for only one experiment; and correcting for any intensity-dependent bias which may be present.

4.3.1.2 Normalisation

Although data plots show a large amount of noise, random error is not such a problem, as there are typically several thousand data points and the degree of correlation between the two channels is generally high. It is the systematic errors which have to be accounted for, and a number of factors have to be taken into account in order to ensure that this happens. For example, variation between arrays is caused by changes in operator, scanner settings, hybridisation conditions, dye bias in different samples and other external factors; while within arrays, the manufacturing process and hybridisation efficiency result in spot intensity variation

over that due to differential expression. In the latter case, unless the lithographic method is used to fabricate the array grid, the quantity of DNA deposited at each spot will vary to some degree and cannot be independently measured. The spot size and shape often also varies randomly or systematically, and it is common for background fluorescence to be variable over the surface of a slide, which will clearly impact on the measurement (if the local background is taken into account when calculating intensities).

Inter-array comparisons are made through array normalisation, which has received much attention over the last few years, due to the significant effect that the choice of a particular method can have on the final expression level data. As with image quantitation, there are several normalisation methods in use, for example the use of 'housekeeping' genes (those whose expression levels do not vary between the two samples), and global subtraction of variously adjusted background figures, scaling of spot intensity distributions, or the more recent and increasingly used rank methods [26]. These methods work well when the expression data are subsequently clustered for similar profiles, but do not take into account any variation based on spot position or intensity.

For studies where the expression levels of individual genes are being examined, replication and more rigorous statistical analysis are necessary, so that significance measures (usually through *t*-tests) can be applied. In these cases, local methods based on regression and statistical analysis are applied, which are very flexible and can show sources of variability which are not visible in the data [27].

4.3.1.3 Interpretation

Corrected microarray data are used to measure quantitative expression levels for particular genes of interest (profiling, with subsequent clustering of behaviour), or to obtain expression profiles over a whole sample, which are then clustered. This reveals trends such as co-regulation, providing clues as to which genes are activated by the same transcription factors, and which are involved in various signalling processes. These studies are clearly dependent on accurate expression levels, and it is vital that all sources of error are explained adequately. The standard clustering approaches used in most microarray software are:

- *Hierarchical* - these methods are used in the generation of phylogenetic trees, and use distance calculations to progressively group data until only one cluster remains, or a stopping rule is applied. These methods are not particularly robust for large data sets, and only offer an overall view of what can be complex similarities. However, hierarchical methods are still widely used, and can provide large-scale viewpoints which can then be investigated further with other methods.
- *K-means* - this approach defines the final number of clusters from the outset, relying on prior biological knowledge to group genes according to a set of properties. K-means (and the related K-medians) clustering methods are prone to multiple stable solutions, depending on starting conditions.
- *Self-organising maps* - this method is a relatively recent addition to microarray analysis [28], and can be used where little information is available about the data. Data points are treated as nodes which are adjusted according to a set of distance rules, and the data set dimensionally reduced.

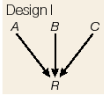
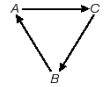
None of these methods has been identified as being a 'best approach', and there are many other, more complex algorithms being developed by academic groups. The well-known issue with all types of cluster analysis is their ability to deliver a variable set of clustered results, regardless of the underlying distribution and even where there is no real relationship between the data.

Significance tests are applied to replicate sets of data where the expression levels of single or small sets of genes are being measured, rather than an expression profile of a whole or composite sample. In this case, standard statistical methods are used to test a hypothesis. As has already been mentioned, microarray data contain large amounts of variation, even after normalisation procedures have been applied, and robust statistical analysis procedures are required to produce useful results. Although the *t*-test is a standard measure of significance, it requires a minimum set of replicates (for arrays, approximately three has been reported as being adequate [29]), and the data set to be *independent* and largely *normally distributed*. Disadvantages include its inherent conservatism and susceptibility to spuriously large *t*-values when the number of replicates is small [30], as is usually the case. Other methods include bootstrapping, which requires large numbers of replicates, and corrections for false positives (which occur when thousands of genes are tested at the same time, even when *p*-values are very small).

4.3.1.4 Experimental design and validation

The cost and resource requirements of microarray experiments place some constraints on the number of replicates available, as well as the experiment/control pairs, and so it is not always possible to work with a data set which contains independent, normally distributed elements. Typically, the design will be hierarchical in that a small number of cell cultures is split among the available arrays, which are then scanned with the red and green channels once or twice each. This produces for each gene a data set which contains dependencies (on cell culture, array and scanning). In these cases, the one or two sample *t*-test is no longer valid, and the analysis software must take account of this by fully separating the sources of variance. Depending on the complexity of the situation, either a standard analysis of variance or a more complex variance components method must be used in order to obtain correct expression and significance levels.

Table 1 | **Single-factor experiments**

Design choices	Number of slides	Units of material (number of samples)	Average variance
Indirect designs			
Design I 	3	$A = B = C = 1$	2.00
Design II 	6	$A = B = C = 2$	1.00
Direct design			
Design III 	3	$A = B = C = 2$	0.67

Variance of estimated effects for three different designs of single-factor experiments. σ^2 was set to 1 throughout.

Figure 5: Approaches to single-factor microarray experiments and their variances. Taken from reference [31]

Yang and Speed [31] have examined this design issue, which has only recently been considered in depth in a few papers, and provide some guidelines on approaches to replication which use resources effectively and minimise variance. Figure 5 shows three different approaches to comparing three samples A, B and C with or without a reference R, the arrays required and the resulting average variance. As these authors state, replication is key to obtaining reliable measures of expression, and should always be used, but the degree of independence must be known and treated accordingly.

Microarray analysis software exists in many forms, as both commercial products and free downloads from academic web sites. For every method of classifying spots and normalising data sets, there is at least one package available, in some cases with full source code. Although comparison studies between some of the common methods have been reported previously [32], these have compared their performance with each other, rather than against reference standards and other experimental techniques. The situation is perhaps aggravated by the fact that there are

currently no standard data formats, although the Axon GenePix file format [22] is widely understood. Oligonucleotide array systems are at present completely proprietary, although the Affymetrix data format is also well known. The MIAME standard (see section 5.3) is addressing this situation, and some journals, most notably *Nature*, now insist that MIAME-compliant data are made available with submitted papers discussing microarray experiments.

Nevertheless, there is a need for the different data analysis methods to be validated against independently measured gene expression on reference samples, so that the most appropriate techniques can be identified for each subsequent method of analysis. This would ultimately lead to a set of 'best practice' methods and fitness for purpose criteria established for different software. In response to this need and to provide a beginning in this direction, a current project funded under the *Measurement for Biotechnology* Programme has been launched with the aim of improving the comparability of gene expression microarray data by assessing the different quantitation and normalisation methods [33]. By assessing the expression levels measured using the most common methods for image quantitation and normalisation (on the same raw data sets) and comparing with a benchmark RT-PCR experiment as well as direct fluorescence measurements, the project aims to identify the key areas of uncertainty and their impact on the different methods.

4.3.2 Proteomics and protein chips

Proteomics - the study of the protein complement of an organism - is increasingly important in that, being concerned with gene products, it provides detailed information about actual cellular processes and protein-protein interactions. Current techniques typically involve the use of 2D separations (isoelectric focusing and gel electrophoresis), followed by further analysis such as mass spectrometry. The slowness of this process, as well as the large number of proteins compared to genes in a typical mammalian genome, means that progress in this area has been less rapid than in genome analysis. Protein chips are an application of the microarray approach described above to increase throughput and resolving power and bring the ability to screen large numbers of proteins rapidly.

Protein chips contain immobilised layers containing agents such as antibodies, which capture target proteins in a highly specific manner. Surface treatment is also necessary to prevent captured proteins denaturing, due to the various biochemical interactions there. For example, the paper of MacBeath and Schreiber [34] describes an early high-density array of proteins, containing 1,600 spots cm^{-2}, and is considered to represent a significant step forward in the possibility of high-throughput proteomics. However, a current limitation on the technology is the number of known specific capture agents available, being numbered only in the thousands at present. Detection methods vary, from those which use labelling techniques to others such as surface plasmon resonance, which do not require any labelling at all.

It is expected that similar issues of comparability will arise when the technology is more established, although many of the sources of variation are related to the stability of proteins on the chip, and the particular methods used to preserve their states of conformation and hydration. These issues lie outside the scope of software metrology, and protein microarray technology itself has to become more robust before a specific study is warranted, but the technology is moving rapidly and the need for such a study will become clear in the relatively near future.

4.3.3 Gene prediction

The large-scale genome sequencing projects, most notably the human and mouse genome programmes, are soon to produce finished physical maps of all the chromosomes, together with complete DNA sequences, consisting of several billion bases per genome. Further sequencing

projects, on other vertebrate species and the relatively little-explored (but extremely large) plant genomes, are also under way. This effort represents an enormous quantity of data to be mined for information about gene location and function. Finding genes in the DNA of a eukaryotic species is a difficult task though, and the scale of the problem can be appreciated by considering that, in the human genome for example, coding DNA accounts for approximately 3% of all bases, that this coding DNA is complicated by the intron/exon structure and splicing, genes often overlap, and that there are many gene-like DNA structures (pseudogenes) which do not produce functional proteins. Two methods have been used to date: sequence comparison with existing genes through database searches, and *ab initio* prediction [35,36].

Comparison methods are based on the likelihood that largely similar sequences are homologous, that is, have common ancestry. Sequence similarity can also suggest functional and structural similarity and, depending on the number and quality of matches returned, information about the function and expression of putative genes can be inferred. These methods are limited by the number and accuracy of protein, cDNA and expressed sequence tag (EST) sequences held in the main curated databases, and structural features such as repeat sequences and ancient conserved regions also complicate matters. Indeed, it is estimated that up to 50% of vertebrate genes have no detectable similarity with already known genes [37]. Consequently, some prediction from unannotated sequences is necessary.

These methods work essentially as a simulation of the central dogma of molecular biology (DNA codes for RNA, which codes for proteins), together with some experimentally observed rules about codon usage, base composition and the characteristics of splicing sites. The basic process is as follows: start and stop codons are located in the DNA sequence, giving a set of potential coding regions. These are translated into protein sequences in all six reading frames, which are then analysed for properties such as GC content and codon-amino acid correlation. The location and sequence of other features such as upstream transcription factor binding sites, splicing sites and polyadenylation recognition sites are also explored, again using the contents of databases for reference. Putative coding regions, with predicted exon regions are then aligned with the main protein and nucleotide sequence databases for existing records with identity above a certain threshold, and decisions made about whether the gene does, in fact, code for that protein. Rogic *et al* [38] have evaluated the major software packages, and conclude that, while prediction accuracies have reached over 90%, there is significant variability over exon length, GC content and species. It is essential to confirm any computational results with experimental techniques such as RT-PCR, and experimental identification of a protein product is the only meaningful way of validating the existence of predicted genes.

4.3.4 Prediction of structure and function from sequence data

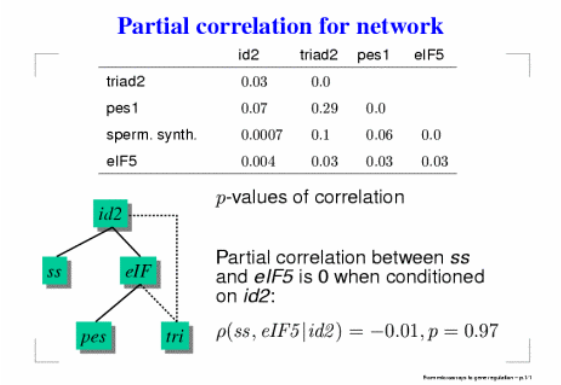
Predicting the three-dimensional structure of a protein from its sequence is one of the central problems of bioinformatics for a number of reasons, not least because the number of solved structures is relatively small (currently some 20,000), and new genes and proteins are continually being identified. In a large proportion of cases, database searches reveal homologous proteins, which can be used as templates upon which to build three-dimensional models. The degree of sequence identity between the query and returned sequences is a measure of the confidence which can be placed in the model. This comparative technique is effective although certain aspects, such as aligning the sequence to the template and modelling loop regions, are difficult. For those proteins which do not have sufficient identity with existing proteins in the central databases for a statistically reliable overall prediction to be made, fold recognition techniques are used. Fold recognition is again a comparative technique, relying on databases of known folds and their sequence patterns, but the relation between sequence and structure is not immediately obvious, and the results tend to be best with single domains.

Those protein sequences for which no homology can be found, for example from new genes identified through gene finding programs or in the laboratory, form a significant though declining group. In these cases, *ab initio* structure prediction has been used for a number of years to obtain estimates of both fold and function, so that database searches for analogous folds may be narrowed, or to provide clues for subsequent laboratory work. Early methods, such as Chou-Fasman [39] and GOR [40], which take a rule-based and statistical approach (that is, using the fact that certain amino acids tend to appear in particular secondary structures, or that features such as reverse turns and binding motifs contain definable patterns), and calculating hydrophobicity values along the sequence, achieved approximately 50% accuracy when tested against known proteins. This has been subsequently improved by the use of techniques such as hidden Markov models and neural networks, and current software can reach over 70% accuracy. More recently, prediction studies have been focused on three-dimensional structure, in which some progress has been made. However, methods based on knowledge of existing structures are more reliable and should always be used where possible.

Software for structure prediction is available as standalone packages and on-line services hosted on the Internet. There are some commercial products in this area, but the majority are from academic groups (for non-commercial use).

4.3.5 Regulatory networks and systems biology

A topic which is rapidly becoming more important with the growth of a detailed knowledge base of gene and protein functions is network and systems biology [41], that is, how signalling pathways, gene regulation and expression and the action of proteins all relate to each other. Microarrays are a powerful tool in identifying phenomena such as co-regulation of gene activity, and gene profiling studies allow correlations to be identified and modelled as networks. These can be Boolean or, allowing for uncertainty, may use probability models (a small example is shown in Figure 6). The aim of systems biology is to relate the wealth of expression and other data to each other in terms of descriptive rules. Such relationships are often complex and nonlinear, requiring advanced computing, such as artificial intelligence, and simulation of dynamic systems. The most advanced project of this type, the Systems Biology Workbench [42], is currently under development.



Source: Birkbeck College Biocomputing Group

Figure 6: One approach to the reconstruction of regulatory networks from microarray data

Systems biology is a relatively new topic, and methods of implementing networks are still very much under development, although there are some tools available which will reconstruct candidate networks from statistical data. Tetrad [43] is an example of one such package which has been applied to regulatory network modelling.

5. Software issues

5.1 Software use

The intention of this report has been to outline the wide variety of applications within bioinformatics which are of importance and how they relate to measurement, and to illustrate the diversity and software dependence of bioinformatics. This dependence covers all aspects, from data acquisition to data mining, although more emphasis is placed on the latter in bioinformatics.

In general, the survey of the subject suggests that the issues surrounding measurement and software in bioinformatics are the same as those present in measurement as a whole, in that the approach to data acquisition and subsequent data processing is the same as it is in other fields of science. Hence, the existing principles of SS*f*M are also appropriate for bioinformatics, where they are applicable to biological measurement.

Bioinformatics uses software of different types, each with its own characteristics and issues, and is heavily dependent upon statistics to produce significant results with the required degree of confidence. It is therefore necessary to ensure that (a) statistical routines which have been programmed from scratch and (b) the underlying methods used in specialist statistical software, use appropriate methods for calculations, taking into account the nature of and dependencies in the data set.

5.2 Software types

There are essentially four different groups of software in use at present:

- Database management systems
- Modelling and prediction software
- Instrument control software
- Statistics software

5.2.1 Database management systems

These include industrial-strength commercial databases such as Oracle and SQL Server, but the use of open-source databases like MySQL and PostgreSQL is also common, particularly in the academic community. Databases are fundamental to bioinformatics, and the database providers recognise this. However, it is data format and exchange, rather than the database software itself, which is the issue. The ability of databases to use formats like XML will aid the process of integrating data from different sources.

5.2.2 Modelling and prediction software

Written primarily in the academic community, but in some cases also available commercially. Results are principally dependent on the quality of data. However, modelling algorithms will often be developed on a case-by-case basis, and while numerical accuracy is not critically important in most cases, design and testing of algorithm performance (in terms of accuracy) is

critical. Algorithms used vary in computational efficiency, depending on whether speed or precision (closeness to the model) is the most important factor.

5.2.3 Instrument control software

This largely applies to instrumentation such as NMR and X-ray crystallography systems, nucleic acid sequencers and array scanners. The first two are long established, with extensive manufacturer documentation in place, and end user organisations tend to have comprehensive validation methods to ensure accuracy. Array scanners are a more recent development but are relatively simple in design, using standard optical data acquisition methods (spot size control, photomultiplier settings, A/D conversion techniques). However, the need for extracting information from large quantities of data can lead to complex implementations. A good example is software for real-time PCR.

In real time PCR, the basic instrumentation includes a laser for fluorescence excitation on a 96-well plate, and a full-spectrum fluorescence detector. The system collects a single excited fluorescence spectrum for each of 96 wells for every thermal cycle (that is, every amplification cycle) - a total of about 4000 separate spectra for each experiment. The software controls the instrument operation; a relatively simple thermal cycling operation followed by a laser fluorescence spectrum acquisition for each of the 96 wells on the plate. But the role of the software extends well beyond the basic control of the instrument and the collection and storage of the spectra. First, the software needs to collect and maintain information about each test sample (typically 16 in total, with six replicates each of calibration and test samples), including any known analyte concentrations. Each time series of spectra for a sample is baseline corrected to remove substantial background fluorescence; the principal spectral components are extracted (probably by multivariate least squares based on the known fluorophore spectra, of which there are three or more per spectrum); the resulting four intensities per cycle (and for each well) plotted against time; a threshold for significant target fluorescence is chosen; cycle numbers estimated by interpolation and the cycle numbers either used to construct calibration curves (for previously identified reference mixtures) or, for test samples, the analyte level estimated from the calibration.

No one of these procedures is especially complex, and the problems for each are well known to instrument and data processing software developers. Yet the integration of all these steps - some essentially database handling, some statistical - into a single semiautomatic process designed for rapid use leads to considerable complexity. Anecdotal evidence suggests that this leads to some flaws in commercial software; for example, one real-time PCR application persistently chose unrealistic cycle time thresholds in early releases, forcing manual threshold selection. However, there is no firm evidence of insoluble difficulty once a problem is recognised.

However, similar problems exist in other fields, and there appear to be no special characteristics in instrument control and data processing that are unique to bioinformatics; the difference, if any, is an increased likelihood of highly integrated control, acquisition and processing, and a substantially higher data throughput.

5.2.4 Statistics software

Many bioinformatics applications, particularly those originating in the academic community, make use of dedicated statistics packages to derive significance levels, ANOVA and other measures. The most widely used of these is R, an open source version of the commercially available S-PLUS. R has been extensively developed and tested, and has an extensive bug tracking project associated with it, and is considered to be reliable and robust. However, R is more of a programming language than a set of statistical routines, and to use it requires some

programming, including the choice of how to implement the built-in routines. Consequently, software issues around statistics are generally concerned with end user implementation of software systems and the production of packages for use and distribution, rather than the software products themselves. Of these, the correct use of the various statistical approaches is key, as many biological data sets are not normally distributed and do not contain independent data points. In fact the tail, rather than the centre, of a distribution is usually the location of interest in bioinformatics, and it is essential to use the correct distribution and statistical approach to calculate reliable *p*-values. This is particularly difficult when measurement data contain discrepant values. Microarrays provide a good example of the importance of this issue: replicates are almost never completely independent, and the number of data points is usually too small to determine the distribution (bootstrapping is often used when dealing with replicate data). However it is worth mentioning here that statistics is a core skill in bioinformatics, and software developers should have the necessary background knowledge to make informed choices in this respect.

5.3 Representation of biological data

Biological data are complex, and present challenges which may be new to the scientist coming from a physical or engineering background. The main aspects relate to the following:

- Terminology
- Data format issues.
- Accuracy and uncertainty information

5.3.1 Terminology

Storing biological data is a complex matter, covering a wide range of concepts and models for the various aspects of living systems. The range of terminology and measurands used in biology is wide, and there is often little consistency between the different specialities, to the extent that certain terms have different meanings in different contexts and variables are of different dimensions and sizes. As bioinformatics projects become more inclusive in terms of the biological problems they seek to address, this uncontrolled use of vocabulary presents an incompatibility issue as serious as those resulting from software and database architectures. There is also a wide range of units, dimensions and scales, all of which must be brought together in a consistent way, and it is a current challenge in bioinformatics to integrate databases of these different kinds of biological information in such a way as to allow mining of data relating to more complete biological systems.

Promising solutions to the problem of specifying the meaning and information content of a biological entity include the development of controlled vocabularies to categorise data in a consistent way. The most advanced initiative in this regard is the Gene Ontology (GO) project [44], which is a consortium-based effort to establish a consistent set of terms - a controlled vocabulary - for the description of gene products in terms of their molecular function, biological processes and cellular components. As such, GO represents a first step towards the ability to construct uniform queries across many databases, rather than unifying them.

5.3.2 Data format and compatibility

To date, there are no standard formats for the biology databases, and in addition to the central, primary repositories for sequences and structures, research groups have produced secondary

databases based on the results of clustering and other grouping processes (for some good examples, see the COGs [45] and CATH [46] databases). Flat text files are the standard at present, but newer technologies are being introduced slowly as consistent formats are proposed, or *de facto* standards appear.

5.3.3 Formats and standards

The well-established biological databases (the main ones being the Protein Data Bank, SWISS-PROT and GenBank) have been in existence for up to thirty years, and have grown exceptionally quickly in recent years. While this growth has resulted in databases with high intrinsic value, it has also brought some problems directly related to storage formats. These are relevant, although not directly linked, to measurement.

At present, most databases in use are of the flat file type, with the records either being retrieved and parsed through Perl scripts and similar methods, or used via web interfaces provided by the curators. Internal layout of records follows a standard form, but the use of text presents several issues which have tended to increase as the databases have grown. These include problems such as:

- *Redundancy* – these databases are not normalised, and contain multiple references to the same piece of information. This introduces the potential for error, as well as the same information being stored in different forms, leading to possible search errors (missed references).
- *Field limitations* – the original design has often not been sufficiently 'future proofed', with data fields not being able to take account of more recent data. A good example is provided by the Protein Data Bank, in which this problem is particularly acute. In this case, measurement techniques and protein science have advanced to the extent that the structures of proteins with large numbers of chains are now routinely solved, while the PDB record format allows only for a small number of chains.

Such problems form a current focus for work which uses technologies such as relational database management systems, which store data in a normalised fashion and allow easy access through structured query language (SQL) queries; and markup languages, which store data and its underlying structure. The most used of the latter is extensible markup language (XML), from which many specialised markup languages for different applications are being derived [47]. In the case of microarrays, which represent an application of major interest to the topic of this report, it is particularly important to standardise on a data format as the knowledge base of data increases, and the technology is used more widely. To this end, MIAME ('minimum information about a microarray experiment') is a set of requirements to be met when submitting data to any public forum, publication or database [48], and its standard implementation is through an XML-based markup language and ontology. This is expected to be in routine use in the near future.

5.3.4 Accuracy

To date, there has been no requirement to provide estimates of the accuracy of data which is deposited in the central sequence and structure databases, the approach being one in which the end user is assumed to know that this is the case, and to treat data accordingly. Nevertheless, with some databases, particularly those containing molecular structures, very detailed experimental information is included with each entry, which provides some indication of resolution. With sequence data, however, it can be difficult to present error information.

Since data quality has a direct influence over at least some of the main bioinformatics applications, this is an issue which may impinge heavily on some applications.

5.4 Software production and validation

5.4.1 Production

Bioinformatics software was, until recently, almost exclusively produced in-house by academic groups and made freely available on the Internet for non-commercial use. This type of software is still the most commonly available, but with the rapid commercialisation of biotechnology and bioinformatics, software firms and instrument manufacturers have begun to offer their own products, based around proprietary methods. Commercial software is generally expensive, costing many thousands of pounds, and as such, is aimed more at the industrial end user, who requires a complete package which includes training and extensive support as well as good performance. However, if one wants to use the most cutting-edge solutions and algorithms, academic software is generally the only source, but the price to be paid is often a poorly thought out user interface and little or no support other than the accompanying literature. Academic software is also usually written by biologists or bioinformaticists who do not have specialist software engineering training, and in terms of maximising performance (as a selling point, or to increase data handling capability) it is generally the view that collaboration of computer scientists and biologists produces the best solutions. However, at least in the initial stages of development, this may not be such an important issue since the purpose of most academic software is to demonstrate the validity of a particular technique, and speed is secondary.

Software error (either in specification or in coding) is of course a feature of both types of software, but it is an arguable point that problems with academic programs tend to be corrected more quickly, as has sometimes been observed with open source software [49].

5.4.2 Validation

Validation studies are well established in the laboratory, through initiatives such as VAM [50], which has specific programmes aimed at addressing topics such as uncertainty in biological measurement; but in bioinformatics they have received less attention. This relates to both the verification of raw data and the testing of the software used, and varies according to research group and software provider.

Some areas, however, have quality assurance practices in place. There is already a useful framework in place for thoroughly testing protein structure prediction software, which can provide a useful guide to fitness for purpose, although at present it takes the form of a competition rather than guidelines. The Critical Assessment of Structure Prediction Methods (CASP) programme [51] is a process open to all developers of prediction software, based at the Lawrence Livermore National Laboratory in the US, and providing a series of real protein sequences which have been solved by NMR or X-ray crystallography for the various groups to derive predicted structures. The assessment stage is very detailed, covering secondary structure assignment, fold recognition and problem areas (for example model refinement, insertions and deletions). Both from-sequence and comparative modelling methods are assessed, as well as ligand-protein interactions ('docking'), and the focus has shifted from secondary to tertiary structure (fold) prediction. This is in line with the continual improvements that have been made in identifying secondary structure sequence patterns.

For fully automated structure prediction methods, such as those hosted on remote web servers, the CAFASP programme [52] has been incorporated into CASP itself, and provides a

framework for evaluating performance without any user intervention. Docking has also been adopted for specific evaluation by the EBI in the form of the CAPRI initiative [53], although this is more a molecular modelling than a bioinformatics problem.

All these initiatives are very competitive, and this has helped to drive innovation and increase the accuracy of the methods. However, they appear to be very much an academic exercise at present - albeit the most rigorous validation procedure there is in bioinformatics, and it is not known if any commercial products are entered.

An important feature of these initiatives is the reliance on test sets. This is likely to be an essential feature of software validation in bioinformatics for some time. Rapid innovation makes it difficult to adopt, develop and test highly robust and portable code; instead, allowing developers to exercise their code on reliable test sets provides consistency of performance testing and improves comparability of output.

6. Measurement uncertainty and statistical issues

6.1 Measurement and traceability in bioinformatics

Before considering measurement uncertainty issues in bioinformatics, it is important to consider the nature of measurement in the field, and the need for metrological traceability*.

Review of the applications of bioinformatics to date (section 4.3) shows a wide variety of measurements. Biological measurements as a whole cover a potentially wider field. The following examples illustrate the variety of measurement types and metrological issues arising.

- DNA profiling or sequencing measures fragment size and identity of terminal base on the way to establishing sequence. The fragment size is estimated by calibration of run length or time against sequence length using 'gene ladders' - well-characterised samples of DNA with a number (usually, today, one per possible base pair) of fragment sizes. The calibrant is typically within an electrophoretic gel run with the ladder interspersed between test samples; more recently, multiplexing allows the calibration to be included in the same channel as the test sample allowing essentially direct comparison. Fragment length, and hence sequence locations, are accordingly traceable primarily to the reference material used. In common with chemical measurement, the material is likely to be well characterised, but not generally certified or traceable to values from a national measurement institute (though in critical applications such as forensic testing, a body such as NIST often supplies reference materials for the purpose). Because of the direct comparison between test and calibration materials, and also the fact that sequence lengths is a discrete (i.e. integer) quantity, uncertainties in position are essentially limited to repeatability issues and can be shown (in forensic profiling, for example) to have negligible effect on identity decisions.
- In array technology, the measurement is a ratio of fluorescence intensities in two test items. One may be (typically is) a reference. But the analyte concentration (indicated by the fluorescence intensity) is not a certified property of the reference; the certified or traceable property is the nature or identity of the material (for example, a specific cell culture, human population, or specific genetic modification event). The relative ratio allows, in current applications, an assessment of the relative degree of gene expression activity under the conditions of test. Measurement uncertainties are dominated by variability from one run to the next. Note, too, that at the current state of the art, the exercise is generally required to show which patterns of expression are consistent and of interest - usually by screening for statistically significant responses and then pattern matching.
- In proteomics, high volume mass spectrometry data can be used to characterise relative protein abundances and hence gene expression and regulation. Few current studies are (because of the low availability of individual reference proteins) capable of accurate quantitation in molar terms, though semiquantitative estimates are possible if similarity

* The term "metrological traceability" is used here to distinguish the concept used in the International Vocabulary of Basic and General Terms in Metrology from the broader concept used more widely in quality systems and regulation. The former refers (ultimately) to the (mathematical) relationship of a measurement result to primary measurement standards which allows a result to be expressed in appropriate units; the latter is wider and refers to the ability to establish the provenance or origin of materials or data. Both are generally important.

of response is assumed. Protein quantitation used for bioinformatics is accordingly largely comparative either within the test material or between test materials.

- In determining the level of genetic modification in Soya or wheat by real-time PCR, quantitation is by calibration on a log scale using a set of standard wheat or Soya flours containing known fractions of genetically modified material and available from the EU's measurement institute at IRMM. The instrument software includes control and all processing and calibration features (this is detailed further in section 5.2, under Instrument Control). Metrological traceability is to the reference materials, and in turn to the mass of genetically modified and unmodified materials used in making up the standards. The genetic composition, however, is traceable (in the wider sense) to the modification event through documentary evidence of material origin and transport.

Note that bioinformatics by no means uses the complete array of biological measurements. Millions of biological measurements - quantitative determination of specific analyte concentrations - are routinely performed each year for individual clinical diagnosis or other reasons, and are unused in bioinformatics applications. Though often challenging, these may be more closely akin to chemical measurements (such as insulin reference material preparation, or cholesterol in blood using GC-MS), though for cost and speed reasons, and because many are specific protein measurements, many use immunoassay techniques which are less common in chemistry. Clinical measurement communities are acutely aware of the need for metrological traceability to improve comparability between measurements, and a great deal of work is ongoing (including that through the BIPM) to establish suitable internationally agreed references and attain metrological traceability. Further, the community has well-established and increasingly sophisticated methods for measurement QA. Yet this community has yet to adopt the metrological terminology and approaches used in physical measurement - particularly with respect to measurement uncertainty. This is partly because, for example, uncertainties in many biological measurements are very large or dominated by biological system variability.

These examples show the striking variability in metrological quality of biological measurement, which clearly covers the range from crude, high-volume comparative measurement through to high quality measurement of individual analyte concentration, calibrated against traceable (or at least, widely accepted) references. However, bioinformatics is currently used chiefly in applications where the measurements are comparative (between test items or with materials of known origin) and do not yet rely heavily on metrological traceability to extract useful information.

6.2 Uncertainty and statistics

6.2.1 Consultation findings

Measurement uncertainty in biological measurement is already the subject of a detailed study within the present M*f*B programme (see below) and detailed additional investigation was accordingly not carried out as part of the present study. This section is accordingly based on interim findings from consultations within the related M*f*B project. Note that these arise in part from a joint workshop, covering both bioinformatics and uncertainty issues in biological measurement, held at LGC in October 2002.

The important statistical issues in this area include substantial dependence on qualitative data (such as pass/fail or identity information), non-normality of distributions and handling of discrete (i.e. integer or otherwise quantised) data. In the field, these are well understood by practising experts in biological measurement, but it is far from clear how such issues are to be integrated into current metrology frameworks. There is also some evidence that analysts at the

bench are less well equipped, especially as these new types of measurement become common in routine analytical laboratories with limited statistical training.

The importance of uncertainty for biological measurement is recognised (see Table 2, which summarises workshop participant responses). The table shows that the perceived importance is high in many cases; in fact, the cases where the most common response is a High importance relate to regulated industrial applications of measurement.

Despite this, as indicated in the discussions related to database applications above, uncertainty information (whether related to the measurement or originating more widely) is rarely included in current databases. Instead, practitioners rely heavily on computed indications of the statistical significance of findings (i.e. on the spread of results associated with particular conclusions) supplemented by their own knowledge of the overall reliability of the data (often a much wider issue than the uncertainties inherent in the measurement results) and on their experience of the reliability of bioinformatics conclusions.

The conclusions relating to uncertainty in that consultation exercise were [54]:

- *"There were no systematic guidelines for the writing of scientific protocols, and this is a key issue that needs to be addressed.*
- *The use of replication, and the estimation of a measurement uncertainty budget for a method, are recognised as being critical in principle. However, on the practical side, the implementation of these is not always feasible, requiring time, money, and expertise. It is often considered better by most laboratories, to spend the time on different, more tangible, outputs.*
- *Analytical science [for biological measurements] is still in its early stages of exploring uncertainty estimates. Because of this, the MfB project may not be able to build on any previous implementations of uncertainty estimates.*
- *Whilst most scientists agree that estimation of method uncertainty is an important contribution in any analytical science, the actual production of an uncertainty budget rarely reaches fruition because of cost, time and expertise considerations. It is generally perceived that these resources would be better spent on more tangible results, for example production of a second set of data to meet customers' requirements.*
- *There is a need for end-users of data to be trained in the terminology and facets of measurement uncertainty, in order for results to be interpreted correctly."*

6.2.2 Activity under the MfB Uncertainties for biological measurement project

It is pertinent to review, briefly, the content and activity of this project here, as it illustrates the state of the art in uncertainty estimation for biological measurement and for bioinformatics. (Detail of the project can be found in the NMS public release document for the M*f*B programme 2002-4).

The work programme is summarised in Appendix 3 for reference. The most important indicators of the state of the art are:

- The case studies are aimed at comparing and contrasting a wide range of measures for uncertainty identification and control, including interlaboratory comparisons, validation etc.
- The study of measurement uncertainty in real-time PCR, using ISO Guide methodology, is believed to be the first so far attempted for a biological measurement related to quantitative DNA measurement.

Thus, the project recognises implicitly that uncertainty of measurement is controlled in biological measurement, but that the metrological concepts and approaches embodied in the ISO Guide to the expression of uncertainty in measurement are essentially absent. The biological measurement community is, because of broadening accreditation, beginning to address these issues, but they are very far from resolved.

The bioinformatics community is a subset of the biological measurement community, and the latter includes far more traceable and quantitative measurements than bioinformatics as currently practised. Bioinformatics is an extremely strong and growing statistical discipline - but recognition, understanding and adoption of metrological concepts of uncertainty are essentially absent among most practitioners. Only in the related measurement communities is measurement uncertainty even recognised as a specific concept.

6.2.3 Uncertainty and statistics - summary

From the discussion above, the most immediate conclusions relating to uncertainty in bioinformatics are:

- Uncertainties in measurement are seen as important principally to prospective regulatory or legislative applications, not to current research or 'lead identification' activities.
- Statistical issues often extend into qualitative and non-normal data distributions, both as a feature of measurement error distribution and population distribution.
- Reliability of the data are compromised by a range of issues, because identification of target analytes, behaviour of micro-organisms and the current very large uncertainties associated with cell biology and biochemistry tend to be much larger than those associated with measurements of relative analyte concentrations. Thus, measurement uncertainty *per se* is a relatively minor issue against a background of much larger reliability problems, and the consequent need is for reasonable control rather than accuracy of estimation of uncertainty.
- While expert practitioners handle uncertainties (in the general sense) in bioinformatics on the basis of experience, awareness of many potential end-users remains low and there are no clear guidelines for conveying or using uncertainty information in applications of bioinformatics.

Table 2: Perceived importance of uncertainty in biological measurement*

Scenarios	Low	Moderate	High	Critical
University/other fundamental research (e.g. seeking disease markers, correlated gene activity etc for detailed study)	X			
Pharmaceutical Research lab seeking new drug targets		X		
Manufacturing facility monitoring biological products for consistency and efficacy			X	
Clinical use which decides patient treatment			X	
Regulators/government screening for possible trends (e.g. in GM use or bio product incidence)			X	
Regulator enforcing limits on bio analyte(s)			X	

*Source: See reference [54]

Key to table:

Low: Would not materially affect outcome
Moderate: Would affect outcome if large
High: Significantly affects interpretation in many cases
Critical: Cannot operate without good knowledge of uncertainty or close control
X points in the table denote the mode of 9 responses.

7. Relevant other NMS programmes

SS*f*M aims principally to support the development of software and (more recently) statistical tools in support of relevant other NMS programmes of work. There are two areas in which biological measurements are immediately relevant; the VAM and the M*f*B programmes.

7.1 Valid Analytical Measurement

The VAM programme [50] is a long established initiative which deals with making analytical measurements more comparable and reliable, including traceability to a recognised standard where appropriate. Essentially, it aims to examine and address sources of systematic error, as well as to help to implement best practice in all aspects of chemical and biological analysis. VAM defines six principles to achieve this:

1. Measurements should be made to satisfy an agreed requirement
2. Measurements should be made using methods and equipment which have been tested and established as fit for purpose
3. Staff making measurements should be qualified and competent to do so
4. Regular, independent assessment of the technical performance of laboratories should be carried out
5. Measurements made in one location should be consistent with those made in another
6. Organisations making measurements should have in place well-defined QA and QC procedures

The VAM programme has historically included some elements relating to biological measurement, and programme formulation at the time of writing includes some biological measurement elements. In particular, the 2003-6 programme is projected to include the following projects (from the Public Consultation Draft);

- *Quantitative DNA Measurements and the Development of Reference Standards*
  (International Standards and Performance Indicators for Nucleic Acid Measurements, A Primary Method to Produce Standards for DNA Quantitation, Development of Standard Units to Measure Gene Expression, Specificity Standards and Reference Indicators for Arrays)

- *Comparability, Quality and Interpretation of Genetic Measurements*
  (Comparability and Consistency of Genetic Measurements, Critical Data Analysis for Low Level DNA Measurement)

- *Advanced Nucleic Acid Technologies*
  (On-Chip Amplification and Process Integration, Whole Genome Amplification, Haplotyping).

Note that although all relate to the improvement of quality in biological measurement, few are critically dependent on software developments. Some develop QA procedures or seek underpinning information to improve measurement quality. Some will seek to make more particularly reliable measurements in order to develop reference standards. In doing so, most will use commercially available equipment and software applied in more rigorous ways or to the

development of measurement standards. In doing so, it is likely that typical laboratory procedures for checking software and equipment suitability and performance will apply. As noted earlier, biological measurement software does not present new kinds of problem; only (in some cases) problems of scale. Existing laboratory validation procedures are accordingly expected to be adequate.

7.2 Measurements for Biotechnology programme

The current M*f*B programme (2001-04) identified five priority themes at the frontier of biotechnology, where there are rapid developments in measurement technology that is critically important in exploiting the biotechnology emerging from the science base, namely (based on published DTI programme information):

- **microarray-based measuremen**t, which is central to product discovery and testing. Building on earlier VAM programme work, this project is intended to address technical issues that represent barriers to comparability between microarray measurements. As earlier discussion in the present report noted, microarray technologies present a range of challenges including some software issues as well as many biological measurement problems.
- **proteomics and genomic**s, the focus of much scientific interest and commercial investment. Proteomics is the subject of significant academic and industrial R&D programmes and it is anticipated that there will be technological breakthroughs and refinements to existing techniques in the short term. The project will identify and address technical barriers with a view to developing comparability in the measurements from the new technologies and in the exchange of the data emerging from these measurements. The long-term aim is to standardise the new measurement technologies and accessibility of the resulting data through measurement method and QA improvement.
- **cell-based testin**g, the way forward in assessing the effectiveness of candidate products, and central to reducing the number of animal tests. The project aims to increase confidence in cell-based testing and to correlate cell-based assay data with genomic and proteomic data in collaboration with industry.
- **physico-chemical methods in biomolecular characterisatio**n, increasingly important in gaining regulatory approval for marketing a bioproduct, and a necessary part of traceable measurement in proteomics and genomics. The project aims to increase confidence in the use of physico-chemical techniques in biomolecular characterisation, by extending the validated limits of established techniques and by evaluating the application of emerging methods.
- **trace biological measurement,** for demonstration of control of contamination, with consideration of the uncertainty of the measurements involved. This project uniquely includes specific elements of statistics and control of measurement uncertainty. It aims to identify current practice in uncertainty estimation and control and to compare the applicability of various approaches to validation, uncertainty estimation and statistical analysis.

The M*f*B programme includes two projects (microarrays and proteomics/genomics) with indirect dependence on software expertise for identification of problems and (in collaboration with industry) their solution. Here, the principal sources of information are the industrial collaborators. It is clear that those involved in developing methods and software are well aware of the problems; anecdotal evidence, however, suggests that not all are fully aware of all the tools available to ensure quality.

In addition, the final project includes elements of statistics and uncertainty estimation. Its principal aim is to characterise and assess those practices currently in place and available. These include routine within-laboratory method validation procedures, between-laboratory studies, uncertainty estimation according to the methodology of the ISO *Guide to the expression of uncertainty in measurement*, and applications of Monte Carlo and bootstrapping techniques.

8. Conclusions and recommendations

8.1 Conclusions

8.1.1 Bioinformatics development

As a field, bioinformatics is highly diverse, gaining greater importance in biology, and developing with great rapidity. As users become more familiar with the fusion of computational techniques with experimental procedure, we can expect to see more papers which describe the use of bioinformatics tools in a highly routine manner.

8.1.2 Data reliability and uncertainty issues

While uncertainty is present and widely acknowledged in bioinformatics data, and the extent of this uncertainty is often unknown, the overall variability of data means that this is not a major issue at present, providing that the correct statistical approaches are used. However, the reliability of the underlying data will be more influential as the sophistication of these tools increases and data mining extracts more information from data. The issue will also become increasingly important where bioinformatics techniques (high volume sample throughput and data handling) are applied in a regulatory context. Thus, uncertainty issues are likely to increase in importance.

8.1.3 Software reliability issues

Software is central to bioinformatics. We have identified four different classes of software, for which the issues differ somewhat because of the state of development:

- Database technology
- Modelling and prediction software
- Instrument control software
- Statistical software

8.1.3.1 Database technology

Database technology is well advanced and offers no fundamental problems within bioinformatics which are not also familiar in many other commercial fields, including those in which data integrity is either commercially vital or a matter of safety. The software itself is accordingly as reliable as can reasonably be expected.

8.1.3.2 Modelling and prediction software

Written primarily in the academic community, but in some cases also available commercially. Algorithms are developed and implemented fast, often, in research areas and perhaps smaller companies, by less experienced programmers. QA depends chiefly on provision and application of reliable test data.

8.1.3.3 Instrument control software

Often complex in order to handle very high volumes of data from many different samples simultaneously, instrument control software is developed case by case with high levels of integration of different functions. This brings risks of programmers implementing algorithms from many different disciplines, and a consequent increased risk of implementation error. This is compounded by time-to-market pressure in a fast-moving field. Awareness of good programming practice for rapid development, together with awareness of a wide range of appropriate, stable algorithms for different purposes, is accordingly important.

8.1.3.4 Statistical software

Like database technology, statistical software is well-established and generally of very high quality, whether commercial or open source/free. However, issues arise in practice due to the need to implement unusual algorithms using standard tools, simply because the algorithms may be poorly understood, or poorly chosen. Testing algorithms developed during R&D activities currently relies heavily on the provision of good test data.

The consistent themes in all these cases are provision of reliable test data, good design practices which encourage application of design and test disciplines, and a need for awareness among a rapidly developing programmer and (sometimes) end-user community. These are recognised issues, and many activities are under way in the bioinformatics and instrument development community. Yet it is not clear how good UK company access is in these fields, and the preponderance of start-ups and SMEs makes it likely that improving access and awareness of existing tools will assist the UK community.

SS*f*M programmes have already encountered similar needs elsewhere, and a range of documentary support is already available. However, it is likely that awareness of these guidelines and tools is low; it is also probable that application in a new discipline will be hampered by terminology issues. Contact with UK companies in bioinformatics - especially in measurement instrument development - is accordingly to be recommended to improve software development practices.

8.1.4 Terminology and data formats

Terminology, vocabulary and data format issues are very substantial in bioinformatics. These issues are well recognised by the bioinformatics community and considerable effort is already under way to address them. Further, many of the issues are concerned with representation within current biological and clinical nomenclature, and must perforce rely most heavily on experts in those fields. SS*f*M activity seems unlikely to be of direct benefit in this context. However, there are relationships between bioinformatics and "data fusion" in other contexts, and there may be opportunities for synergy.

8.1.5 Relevant other NMS programmes

The two principal programmes which might benefit from SS*f*M support or collaboration are the VAM (DNA measurement) and M*f*B programmes. However, there are few areas of SS*f*M interest which are unique to biological measurement, and numerical accuracy and even software reliability are rarely critical provided that obvious pitfalls are identified and eliminated. The most urgent needs are accordingly to bring biological measurement practices in general up to those in chemical and other measurement fields.

In the longer term, it is possible that biological measurements will begin to develop approaches and protocols related to uncertainty estimation following ISO guidelines. Where this is the case,

it will clearly be important to address the often asymmetric and discrete distributions encountered. These issues are, however, already being addressed within SS*f*M, and provided that they continue to receive support, will be applicable when the need arises.

8.1.6 Relevance of SSfM

As we have stated, the existing principles of SS*f*M relate well to bioinformatics software, and the bioinformatics community (academics, commercial software developers and end users) should be made aware of and encouraged to participate in SS*f*M activity. This is advantageous in that those issues identified as being particularly important in biology (data compatibility, terminology, and integration of diverse data types) would then be included, giving more complete coverage of software and data issues in science as a whole. Specific SS*f*M activities that are relevant include:

- *Data fusion*: Issues in data fusion in other fields should map well onto bioinformatics data interpretation issues
- *Test set generation*: SS*f*M's methods for producing reliable data sets for statistical software could provide a useful basis for some type of bioinformatics software validation.
- *Instrument software development and validation*: Existing SS*f*M guidelines and practices could be applied in biological measurement instrument development and application.

Note that SS*f*M activities might also be expected to benefit from knowledge gained by the bioinformatics community. Applications such as data mining and statistical handling of non-normal distributions and extreme values in large data sets may see applications in other areas of SS*f*M interest. Thus it is sensible for SS*f*M to review and test tools emerging from the bioinformatics area, particularly where they might apply to metrological development.

8.1.7 Technology

In identifying areas for SS*f*M activity, it is important to identify the appropriate technical fields. It will be clear from this report that the one area which has progressed sufficiently from the research laboratory to wide acceptance in industry is microarray technology. This technology integrates measurement, data processing and interpretation of very high data volumes to an unusual degree.

Although protein arrays are fast being developed, the technology is at a much earlier stage than in DNA array technology, and both the technical difficulties associated with the immobilisation and stabilisation of protein microspots, and the specificity of protein array applications (which will not help the development of open and standard formats), mean that DNA microarrays are unlikely to be superseded in the foreseeable future.

It is accordingly considered most appropriate that SS*f*M activity in bioinformatics should be directed into support for DNA array technology implementation in the near term.

8.2 Recommendations

The key themes arising in this report suggest that SS*f*M activities are already well applicable to biological measurement. However, it is important to be aware that a wider range of statistical procedures is in use than in many measurement areas, that awareness of uncertainty issues is generally far lower, and that measurement instrument control software tends to be both complex and rapidly developed.

These conclusions suggest the following general recommendations in developing SS*f*M:

- Assess and, if necessary, act to improve the quality and time-to-market of biological measurement software provided by UK companies. This should involve a sector-specific review of the current state of the art among the UKs bioinformatics measurement software developers which reviews the software specification and development process; algorithm use and implementation; appropriateness of algorithms; numerical accuracy; and generation and use of test data. The review should make recommendations for provision of guidance, software tools and training.
- Make provision for communication on measurement uncertainty issues at least with those UK measurement institutes involved in measurement standards provision and preferably with bioinformatics community representatives, to identify specific needs for guidance and to address those needs in collaboration with other relevant NMS programmes.
- Publicise SS*f*M's activities on reference data sets with a view to a) stimulating the adoption of similar principles for reliable bioinformatics test set development and b) in the longer term, identifying and supporting test set development projects. For bioinformatics, this can be best achieved through development and application of pilot test sets for software intended for statistical analysis of large data sets, including sets with discrepant values.
- Review statistical tools (such as data mining tools) emerging from bioinformatics research to assess their applicability to other measurement fields and to metrological development, and conduct feasibility studies on prospective applications.
- It is recommended that industrial contacts focus in the first instance on DNA microarray technology, as this is the best developed field and a good model for future interaction.

In addition, the relatively low level of response obtained during this study indicates that further consultation is advisable. It is accordingly recommended that this report and its recommendations be made available to organisations who have not contributed - and particularly to NIBSC - to stimulate additional input to the SS*f*M formulation process.

9. References

1. Quote taken from www.bioinformatics.org
2. *NIH Working Definition of Bioinformatics and Computational Biology* (2000), at www.bisti.nih.gov/CompuBioDef.pdf
3. NM Luscombe, D Greenbaum, M Gerstein, *Method Inform Med*, 4/2001, 346
4. B Rybak, *Psyché Soma Germen*, Gallimard, Paris (1968)
5. www.rcsb.org/pdb
6. www.ncbi.nlm.nih.gov/Genbank/index.html
7. ES Lander *et al*, *Nature*, **409** (2001) 860
8. JC Venter *et al*, *Science*, **291** (2001) 1304
9. www.ensembl.org/Mus_musculus
10. Figures taken from www.abpi.org.uk
11. Peter Goodfellow, Director and Senior VP Discovery Worldwide at GlaxoSmithKline, at *Microsystems, The Big Future*, a new technology conference hosted by DTI (April 2001). Further details are available from the author
12. Kurt Schilling of the Foresight Team at Unilever plc, at the meeting cited in [11]. The example given was of gene expression effects in the skin as a result of exposure to the sun, and how new treatments might combat them
13. www.celera.com
14. S Schweigert, P Herde, R Sibbald, *Computer Applications in the Biosciences* (now *Bioinformatics*), **11**,4 (1995) 339
15. *Medical Informatics: Computer Applications in Health Care*, EH Shortliffe, LE Perreault, G Wiederhold, LM Fagan (eds), Addison-Wesley (1990)
16. Details of the PHRED (base calling) and PHRAP (sequence assembly) programs can be found at www.phrap.org
17. HH Otu, K Sayood, *Bioinformatics*, **19** (2003) 22
18. J Knight, *Nature*, **410** (2001) 861
19. *Microarray Biochip Technology*, M Schena (ed), Eaton (2000)
20. See www.affymetrix.com or www.ogt.co.uk for details of the oligonucleotide array method
21. ScanAlyze, a microarray image analysis program available from the Eisen Lab (rana.lbl.gov/EisenSoftware.htm), uses this method
22. For example the GenePix suite from Axon Instruments
23. R Adams, L Bischof, *IEEE Trans Pattern Anal Machine Int*, **16** (1994) 641
24. YH Yang, MJ Buckley, S Dudoit, TP Speed, *Berkeley Technical Report #584* (2000). Available at stat-www.berkeley.edu/tech-reports
25. C Kooperberg, TG Fazzio, T Tsukiyama, *J Comp Biol*, **9** (2002) 55

26. E Schadt, *J Cellular Biochem* (supplement), **37** (2001) 120
27. YH Yang, S Dudoit, P Luu, TP Speed, *Berkeley Technical Report #589* (2001). Available at stat-www.berkeley.edu/tech-reports
28. P Tamayo, D Slonim, J Mesirov, Q Zhu, S Kitareewan, E Dmitrovsky, ES Lander, TR Golub, *Proc Natl Acad Sci USA*, **96** (1999) 2907
29. ML Lee, KC Kuo, GA Whitmore, J Sklar, *Proc Natl Acad Sci USA*, **97** (2000) 9834
30. W Pan, *Bioinformatics*, **18** (2002) 546
31. YH Yang, T Speed, *Nature Genetics*, **3** (2002) 579
32. TC Kroll, S Wolfl, *Nucleic Acids Research*, **30** (2002) e50
33. *Comparability of gene expression measurements on microarrays*. This is an LGC-led project funded under Theme 1 of the *Measurement for Biotechnology* programme, and has the aim of establishing guidelines for ensuring comparability of microarray data by examining the most common normalisation techniques currently in use. See www.bioindustry.org for details or contact the author
34. G MacBeath, SL Schreiber, *Science*, **289** (2000) 1760
35. C Burge, S Karlin, *J Mol Biol*, **268** (1997) 78
36. compbio.ornl.gov/Grail-1.3
37. JM Claverie, *Hum Mol Genetics*, **6**,10 (1997) 1735
38. S Rogic, AK Mackworth, FBF Ouellette, *Genome Research*, **11** (2001) 817
39. PY Chou, GD Fasman, *Biochem*, **13** (1974) 222
40. J Garnier, DJ Osguthorpe, B Robson, *J Mol Biol*, **120** (1978) 97
41. H Kitano, *Science*, **295** (2002) 1662
42. www.cds.caltech.edu/erato
43. The Tetrad home page is at www.phil.cmu.edu/projects/tetrad
44. www.geneontology.org
45. www.ncbi.nlm.nih.gov/COG
46. www.biochem.ucl.ac.uk/bsm/cath
47. www.cellml.org/public/documentation/biological_mls.html
48. www.mged.org/Workgroups/MIAME/miame.html
49. The Open Source movement (for examples, see www.opensource.org and www.fsf.org) has long claimed that shared source code and free distribution results in much faster development and bug resolution than closed, proprietary products
50. See www.vam.org.uk
51. The CASP web site is at predictioncenter.llnl.gov
52. www.cs.bgu.ac.il/~dfischer/CAFASP3
53. capri.ebi.ac.uk
54. Source: Minutes of meeting, “Issues in Measurement Uncertainty”, 10th October, 2002

Appendix 1: Organisations contacted

The following organisations were approached during the course of this study. Approaches were limited to individuals or organisations with existing contact with LGC or NPL who had previously agreed to discuss biological measurement issues.

BioVex.

Birkbeck College, School of Crystallography*

Blood Products Limited

European Bioinformatics Institute*

GlaxoSmithKline

Imperial College, Centre for Bioinformatics

LGC Limited*

National Blood Service.

National Cancer Research Institute*

National Institute for Biological Standards and Control

NPL*

OxaGen

Oxford BioMedica

Oxford GlycoSciences

PHLS Colindale

PHLS Newcastle

PHLS Preston

Rank Hovis McDougall*.

Renovo*

Sanger Institute

Severn Trent water.

Thames Water

Unilever*

University of Manchester

*Involved directly in meetings/discussion.

Appendix 2: Focus group meeting summary

This meeting was held at LGC in October 2002, and covered both uncertainty in biological measurement in general, and software/measurement issues in bioinformatics. The details have been reported separately, but are summarised here. Attendance was as follows:

Malcolm Burns	LGC
Maurice Cox	NPL
Steve Ellison	LGC
Carole Foy	LGC
Sarantis Kamvissis	NPL
Jacquie Keer	LGC
Magdalena Sara	National Cancer Research Institute
Swen Tromm	European Bioinformatics Institute
Gordon Wiseman	Rank Hovis McDougall

After an overview was given of the SS*f*M themes and an illustration of the key software issues in bioinformatics, the discussion brought out the following points:

- Large data sets (over 10,000 points) are frequently difficult to handle, in cases where every data point is important, and manual quality pre-screening is necessary. Systematic detection of outliers is also a problem, and in some instances it is the outliers which are important (in fact, in bioinformatics it is the tails of distributions which are usually important). End user expertise is thus critical for screening and inspection.
- Estimates of experimental error are sometimes included in the sequence and structure databases, but the true uncertainty levels are usually not known. Treatment of uncertainty is usually left to the end user of the data, but since some applications are uncertainty-critical, some reliability measure is needed.
- A simple example would be database searching using BLAST, where clearly the quality of sequence data and how uncertainty is dealt with potentially affect the number of hits returned. However this is open to question, and it is not known to what extent measurement uncertainties affect the final conclusion (in microarrays, the user accepts noisy data, pre-screens and takes this into account). Different application areas also have different requirements for data interpretation (e.g. healthcare vs. research).
- Data reduction techniques (use of metadata) are often necessary for visualising large data sets, such as crystallography or sequence data. There are currently no rules on quality control measures or naming, and again filtering is left to the user. In some areas, 'cleaner' data sets are required.
- The need for replication and experimental verification of results was also highlighted.

Appendix 3: Unilever Meeting report

A meeting was held with the statistics department at Unilever Research Port Sunlight in September 2002, to discuss the main issues in bioinformatics from the viewpoint of the consumer products industry. Those taking part were:

Simon Cowen	LGC
Steve Ellison	LGC
Mojgan Naeeni	Unilever, statistics dept
Ken Rabone	Unilever, statistics dept

Representatives from other groups at Unilever were also present. The discussion began with a presentation of an application of current interest, and one which is typical of the data handling requirements of the consumer products industry. Unilever is applying alternative statistical methods, in this case logistic regression, to the testing of new disinfectant formulations. As is well known, logistic regression is a technique which converts a binary response (a positive/negative) to a continuous variable ranging between zero and unity, thus allowing screening (does it work or not?) to be expressed in terms of a probability. This is a method widely used in bioinformatics.

The aim is to determine the minimum dosage required to prevent bacterial growth and satisfy performance requirements under varying conditions. It is amenable to high-throughput screening because a high-density plate containing different formulations can be scanned quickly, and the binary data obtained then converted into probabilities once a suitable threshold has been set. The software and data visualisation tools used are the same as those used for microarrays (with which this application mirrors in a number of ways), and Unilever are currently using Spotfire, a commonly used commercial package. Data points of particular interest are those which lie at the boundary between the bacterial growth and non-growth regions, and measures of statistical significance are important. Unilever intends to validate this approach and use it as a standard method; to do this, it is necessary to quantify the variability and reliability of the predictions made by the method, that is, how trade-offs between the variables (concentration, formulation) affect results. Other issues raised as being important, included:

- *Measurement* - replication was stated as being essential, and it is important that measurement software and procedures apply meaningful tests of significance. In some cases, few replicates will suffice, but an appropriate minimum number should be chosen. Decisions are currently analyst-driven, rather than automated.
- *Data reliability* - areas where the reliability of data and the results of data processing are key for consumer products include metabolomics, microarrays and proteomics, which are all concerned with the interaction between the body (mainly skin) and various formulations. This need for accuracy and knowledge of measurement uncertainty extends even to PCR products and nucleic acid sequences.
- *Data sets* - these are frequently large and multidimensional, and include sets of data on consumers with parameters such as preferences and other behaviours.
- *Data format and storage* - the industry is beginning to move towards the use of electronic notebooks, with its easy incorporation into audit trailing. The need to comply with best practice systems such as GLP is very important to the industrial user, particularly those which are heavily regulated, and open, hierarchical data formats are necessary to do this easily

Appendix 4: M*f*B Uncertainty Project work programme

The work programme is structured into five work packages:

Work Package 1. Identification of current practice in assessing and controlling uncertainties

A paper and consultation-based study of current practice in academia, large companies and SMEs involved in biological R&D and measurement. We propose a two-staged approach, mixing focus group identification of the main issues with more detailed follow-up. This study will identify the range of current approaches and inform the selection of case studies for experimental study.

Work Package 2. Communication and dissemination

Dissemination is provided for through a specific work package. The work package provides for direct contact with regulators, accreditation bodies and the measurement community through a range of targeted and broadcast activities (given in detail under 'deliverables' below). These measures will both inform a wide audience and reach key decision makers involved in regulating and controlling good measurement practice in the sector.

Work Package 3. Interlaboratory studies for uncertainty assessment and control

Interlaboratory studies are a crucial tool in the development of widely accepted standard methods of analysis and testing. Ordinary statistical tools for handling interlaboratory study data for quantitative methods are of limited value for the frequently non-normal or qualitative data generated by many biological tests. Using two case studies, this work package will illustrate the issue and demonstrate alternative methods for treating the data.

Work Package 4. Single-laboratory validation and QC approaches

Method validation is a powerful and effective experimental method of identifying factors affecting chemical measurement and testing methods, verifying assumptions used and demonstrating that method performance meets requirements. Generally similar approaches are used in measurements for biotechnology, but require adaptation, again, to the different problems facing biological measurements. Two case studies will demonstrate practical approaches that work in ordinary testing laboratories. The principal outcome of the work will be a report on best practice in single-laboratory validation for biological measurement, supported by examples from the case studies.

Work Package 5. Formal uncertainty budget approach

Uncertainty budgets based on formal mathematical models are, so far, rare in biological measurement. Reasons include the qualitative nature of many measurements, and large and poorly understood effects such as trace inhibitor effects. Yet the potential of the methodology for improving measurements, well demonstrated in other sectors, suggests that promotion of the methodology will bring benefits fir some measurements, particularly where quantitation is important. A detailed case study of a relevant application with potential for clear benefit is accordingly proposed. The case study suggested is based on a specific real-time PCR method likely to be of considerable importance for GMO regulation among other applications.