Biomolecular Function and Evolution in the Context of the Genome Project

(July to December 1998)

Report from the Organisers: PJ Donnelly (Oxford); W Fitch (Irvine); N Goldman (Genetics, Cambridge)

Introduction

There is a long and productive history of interplay between genetics on the one hand and mathematics and statistics on the other. The "molecular revolution" over the last 15 years, and in particular the impetus of genome projects, has transformed the field into one with an abundance of data and a paucity of relevant mathematical models and techniques. By 1998, the maturation of genome projects had made data on DNA, proteins, gene duplications and gene arrangements on the chromosomes widely available. These data will have a profound impact on the practice of biological research, and, ultimately, medical diagnostics and preventive medicine.

The aim of this programme was to bring together world-leading researchers in molecular biology, biological mathematics and computer science to meet and collaborate for extended periods on bioinformatic problems arising from the analysis of the current flood of molecular genetic sequences and structures. These include topics on subjects such as probabilistic modelling, statistical data analysis, stochastic processes, geometry, computational

complexity, neural networks, genetic algorithms and expert systems. These topics were particularly apt for a Newton Institute programme, with the UK being a world-leader in molecular biology, but being generally less well-developed in the application of mathematics to the resulting data analysis problems.

Many challenging biomathematical research topics were raised and, as a consequence of recent advances in computational statistics, vast improvements in the quality of statistical analyses of these data were shown to be possible and various collaborations to that end initiated. Earlier parts of the five-month programme concentrated on explicitly evolutionary topics such as phylogenetic trees and networks, comparative analysis using evolutionary trees, population genetics topics and viral evolution, with conferences on human evolution and viral evolution. The later parts of the programme concentrated on structural, functional and genomic topics, with time devoted to secondary and tertiary structure prediction, fold recognition, motif and pattern recognition, hidden Markov models and gene prediction. Below, we report on the activities of the programme in the order they occurred, generally a week at a time.

Organisation

The overall organisation was undertaken by Peter Donnelly (Oxford), Walter Fitch (UC Irvine) and Nick Goldman (Cambridge). A semi-formal arrangement had the scientific programme co-ordinated by Walter Fitch in July and August, by Peter Donnelly in September, and by Nick Goldman from October to December. Day-to-day administration of the programme was carried out by Nick Goldman.

A number of the programme participants played important roles in the organisation of workshops and theme weeks, and many of these contributions are mentioned below. In addition, the organisers want to note the efforts of David Balding (Reading) as "social events co-ordinator" throughout the programme. Three seminars in the Newton Institute Seminar Series were presented by programme participants: *Predicting the future evolution of the human influenza virus* (Walter Fitch), *Two problems of multiple comparisons in molecular genetics* (David Siegmund, Stanford) and *Experimental and computational approaches to analysing DNA-protein interactions* (Gary Stormo, Colorado). The programme also hosted the XXI Fisher Memorial Lecture, *Mathematics of genetic diversity before and after DNA*, presented by Prof Sir John Kingman and chaired by Prof Sir Walter Bodmer, and the annual one-day Summer Outing of the British Region of the International Biometric Society, at which Biometric Society members were shown around the Institute and heard three lectures by Institute programme participants, including one contributed by our sister programme.

Participation

The programme attracted 69 long-term participants (average stay approximately 7 weeks), and 126 short-term participants. A very high proportion of the world's leading researchers was able to attend the programme at some stage, although the nature of many molecular biologists' laboratory work meant that many were unable to attend for as long as they or we would have liked. Despite this, the organisers worked hard to ensure that participants were at all times drawn from the world's top scientists and not simply from the world's most mobile scientists.

The UK is a world leader in experimental molecular biology, but is less well-developed in the application of mathematics and statistics to analysing and understanding molecular genetic data. At all times, we attempted to co-ordinate the presence of leading researchers from around the world with the attendance of the UK's leading and most promising young

researchers. Approximately 65% of the programme's core funding was used to support UK researchers.

Scientific Programme

In order to create the maximum possible level of continuity as the programme progressed, we decided to split the programme into 'theme weeks', the scientific content of each of which would be organised by one of the programme organisers and one or two co-opted programme participants (listed below). Theme weeks typically contained from 5 to 10 talks, both prearranged and also impromptu, given by invited speakers and other attendees. The theme weeks were interspersed with workshops, on topics related to the surrounding themes, and a number of free weeks in which participants were entirely at liberty to pursue their own interests. Overall, there was a broad divide into evolutionary topics for the first half of the programme, and genomic topics for the second half.

Evolution: Phylogenetics

(20 July to 30 August)

One of the first goals in bioinformatics is frequently to discover the historical relations among the sequences one examines. There are several steps in this process, the first of which is to align one's sequences. This is followed by finding the best fitting tree and that by an attempt to understand other biological phenomena in the context of that phylogeny.

Alignment (W Fitch; T Smith, Boston)

The study of molecules in an evolutionary context requires one first to have a set of homologous sequences (nucleotide or amino acid), where homologous means that they have a common ancestor. It is then necessary to align these sequences, one under another, so that not only do the sequences possess a common ancestor, but all the nucleotides (amino acids) in a single column have a common ancestral nucleotide (amino acid). As that ancestor becomes more and more remote in time, it becomes more and more difficult to discern which nucleotide belongs in which column. The first week of the programme, appropriately, was devoted to this primary topic. Topics covered included both optimal global and local alignments and such important details as the effect of different nucleotide frequencies in the sequence and unequal rates of evolution along the sequence. Also considered were approaches that obtain the alignment even while undertaking to obtain the phylogeny at the same time.

Phylogeny (W Fitch; D Penny, Massey)

The next step is to use the aligned sequences to obtain a phylogeny, a tree of ancestral relations. The second 'Evolution week' included scheduled presentations including talks about the myths that already inhabit the discipline, the effects on methods which were designed for small numbers of sequences when these numbers are scaled up by several orders of magnitude, the problem of multiple nearly equally good trees and the use of networks rather than trees to represent phylogenetic relations.

Workshop: EC Summer School (W Fitch; J Felsenstein, Seattle)

This week was an interruption of the progress on deep problems, and capitalised on the presence of many (indeed, most) leading molecular phylogenetics theorists and public domain software developers to provide a hands-on experience for novices in the use and potential of many of the programs for analysing data. The Summer School, entitled *Methods for Molecular Phylogenies*, was held from 10 to 14 August.

Many methods in molecular phylogenetics were devised in an ad hoc manner by biologists not trained in statistical methods. Consequently the field became controversial, with ill-understood and 'opposing' methodologies being introduced. Only in the past 10 years has a

fuller understanding of the statistical properties of phylogenetic inference methods enabled a truly scientific framework for data analysis to be developed. A large proportion of established researchers around the world are still not fully aware of the 'state of the art' in molecular phylogenetics. We felt it was valuable to devise a course to introduce younger scientists to modern ideas on the major methods of data analysis, the mathematical and statistical foundations of these methods, and, in a practical vein, the use of the major computer programs available for performing these analyses.

The course very successfully met these aims, through its daily mixture of lectures (2 to 3 hours each morning) and computer-based practical classes (afternoons). Computer facilities were kindly provided by the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, and students typically worked with one pair per computer with various opportunities for individual work. The topics covered included the popular methods of parsimony, distance methods and maximum likelihood; optimal searching strategies of tree space; statistical tests in phylogenetics; and advanced uses of phylogenies. The course was attended by 73 official participants, including 51 from 13 EC countries; in addition we were delighted that up to 30 other scientists from Cambridge and elsewhere attended the morning lecture sessions. We had many more wanting to attend than we could admit and the reception was very positive. Our and others' experience leads us to believe that this indicates a great need in Europe for such a course to be given annually.

The Summer School's scientific content was organised by Joe Felsenstein (Seattle), and the computer sessions were co-ordinated by Frank Wright (BioSS, Dundee). Other lecturers were: AWF Edwards (Cambridge), W Fitch, A Rzhetsky (Columbia), J Huelsenbeck (Rochester), M Charleston (Oxford), A von Haeseler (Munich), M Newton (Madison), Z Yang (London), J Hein (Aarhus), W Maddison (Arizona), R Page (Glasgow).

Beyond Phylogenies (W Fitch; P Harvey, Oxford)

Once one has a tree of relationships, what can one do with it? This is what the fourth Evolution week was about. Topics included were correlates of species richness, animal competition, speciation, ecological diversity, estimating ancestral character states, and macroevolution.

Evolution: Population Genetics

(31 August to 4 October)

Population genetics is concerned with studying the diversity observed at the DNA sequence level between individuals in a population. Part of this, the collection of data, is clearly empirical. The correct interpretation of such data poses challenging mathematical and statistical problems. The data represents an incomplete snapshot, taken at a single point in time, from the evolution of the population. Such data is typically high dimensional, with a complicated correlation structure arising from the extent of shared ancestry between the chromosomal regions sampled. There is a well-developed tradition of mathematical modelling, typically in a stochastic way, of the evolutionary processes. These models provide insight into the ways in which observed patterns of variability depend on both the genetic factors at work (mutation, selection, recombination) and the demographic history of the population. Less well developed are statistical methods for inference from such data.

***Population Genetics** (I)* (P Donnelly; R Harding, Oxford)

The week aimed to set the scene for the focus on population genetics by bringing together leaders from both the experimental and modelling side, with a view to looking forward at the types of data which will be becoming available and the consequent challenges for developing appropriate models and methods of analysis. Both through formal presentations and structured discussions, and informal interactions, it appeared to succeed.

Workshop: NATO ASI: Genes, Fossils and Behaviour
(P Donnelly; R Foley, Cambridge; S Paabo, Munich; A Rogers, Salt Lake City)
One particular organism of considerable interest is humans. Studies of human history have been seriously undertaken for at least 100 years. The molecular genetic revolution has added a new tool to the existing armoury (fossils, archaeology, language, behaviour, climate), since patterns of extant human genetic diversity have been shaped by the patterns of human demographic history. The ASI aimed to bring together leading researchers from across the range of disciplines involved, to give state-of-the-art lectures to young workers, and to try to integrate much of the existing disparate work. It succeeded beyond even the optimistic hopes of the organisers, in no small part because of the constraints imposed by NATO (limited numbers of 60-minute lectures each of which was followed by 30 minutes of timetabled discussion, a meeting of at least 10 days, a focus on bringing younger workers).
The invited lecturers were Paabo (Munich), Donnelly (Oxford), Foley (Cambridge), Lahr (Sao Paulo), Ward (Oxford), Bertranpetit (Barcelona), Barbujani (Milan), Hublin (Paris), Griffiths (Monash), Takahata (Tokyo), Jorde (Salt Lake City), Stoneking (Penn State).
***Population Genetics* (II)** (P Donnelly and S Tavare, Los Angeles)
The week focused on the specialised area of computationally intensive likelihood-based methods for inference from molecular population genetics data. This represents a new subarea of the field (the first papers are from about 5 years ago) which is extremely promising. Representatives of all the leading groups were present. While the specialisation excluded other programme participants more than for other weeks, there are obvious advantages to having high-level discussions amongst leaders of the field, and those involved were positive about the
experience.
Workshop: Viral Evolution (N Goldman; E Holmes, Oxford; A Rodrigo, Seattle)
The aim of this workshop, held from 5 to 9 October, was to foster interactions and collaborations among virologists, evolutionary biologists and mathematicians, to discuss how best to analyse the increasing amounts of nucleotide sequence data being obtained from viral genomes, and to address the evolutionary and functional implications of the data already available. This data resource, coupled with the desire to control the spread of viruses through human populations, makes the study of viral genomes one in which an integration of data and theory is very likely to be profitable, both to the medical and biological sciences and also to offer inspiration to theoreticians. The study of viral genomes is an area of biological research that has often witnessed the successful introduction of new methods of data analysis.
Within this general framework, a wide variety of topics was covered. Workshop sessions were devoted to Case Studies in Viral Evolution, Rates of Viral Evolution, Within-Host Evolution of HIV, Evolution of Drug Resistance in HIV, Viral Adaptation, Viral Quasispecies and New Uses for Viral Phylogenies. The workshop was deemed to be highly successful by all those who participated (approximately 50 in number). In particular, the meeting whole-heartedly followed the ethos of the Newton Institute in that discussion time was given special emphasis.
The workshop's scientific content was organised by Edward Holmes (Oxford) and Allen Rodrigo (Seattle). Other speakers were: T Gojobori (Mishima), H-U Bernard (Singapore), W Fitch, J Drake (North Carolina), C Bangham (London), P Simmonds (Edinburgh), S Frost (Edinburgh), Y-X Fu (Houston), P Zanotto (Sao Paolo), J Albert (Stockholm), K Crandall (Provo), S Bonhoeffer (Oxford), J Brookfield (Nottingham), L Chao (Maryland), A Sasaki (Kyushu), H Bourhy (Paris), D Smith (Edinburgh), T Leitner (Stockholm), P Sharp (Nottingham), M Pagel (Oxford).

Genomics: Protein Structure

(19 October to 8 November)

One of the major, and largely unsolved, problems in bioinformatics is the prediction of protein structures from their DNA or amino acid sequences. Sequences are relatively easy and cheap to determine, whereas the experimental determination of the encoded protein's structure, for example by X-ray crystallography, remains difficult, time-consuming and expensive. Yet it is a protein's sequence that determines its structure which, in turn, determines its function and so holds one of the keys to understanding all forms of life. The implications for scientific research, medical applications and commercial exploitation of reliable methods to infer protein structures from their sequences are enormous, and attract continued interest and investment. Three consecutive theme weeks of the programme (19 October to 6 November) were devoted to analysing the current state of affairs in protein structure analysis, concentrating respectively on methods based on sequence analysis, on comparison with known structures, and on the combination of these approaches.

Sequence Analysis (N Goldman; G Barton, Hinxton;
W Taylor, London)

In some senses, the 'Holy Grail' of protein sequence analysis is the inference of protein function from protein sequence. In practice, this is found to be too difficult even to attempt and the problem is split into simpler pieces, each perceived to be approachable (although all are hard and as-yet unsolved). Starting with protein sequences, what is considered manageable is the prediction of protein secondary structure. Proteins are almost invariably composed of 'building block' structures which are joined together in specific manners to create the three-dimensional structure of the complete protein. The number of building block structures is very small (around a dozen, of which three predominate). The secondary structure of a protein is the categorisation of its constituent amino acids into these common elements. Although the 3-D structure of proteins (ie the determination of the co-ordinates in space of each of the thousands of atoms of a protein) is not currently estimable from sequence data, it is widely felt that the secondary structure problem is both possible, and would be of great value towards the inference of first 3-D structure and thence function. The first of these three theme weeks was aimed at reviewing the state of the art in protein secondary structure prediction.

Structural Analysis (N Goldman, A Lesk, Cambridge, W Taylor, London)

The second 'Protein Structure' theme week was devoted to reviewing the state of the art in our understanding of protein structures. This can be viewed as the other half of the problem outlined above; given what we already know about protein function and the structures that perform those functions, how can we curate our knowledge and find important patterns that might link to the inferences that we can make directly from sequences? One important topic discussed during this week was the relationship between, and strengths and weaknesses of, the various classifications of protein structures that currently exist. A number of participants found this a valuable exercise, particularly since it took place in front of the scientists who had generated the classifications. This very much clarified the situation and showed quite precisely what future steps should be taken both to reconcile the differences and to interpret the results arising from their use. Another recurrent theme was the surprisingly low number of fundamentally different protein structures that have been found to date. Early theoretical studies suggested that the range of protein structures that could in principle be formed by chains of amino acids was huge; in practice, it is found that there is very considerable duplication of structures and that proteins with very different evolutionary histories and functions tend to fall into the same structural 'families' or 'superfamilies'. There were valuable exchanges between those who maintain classifications of protein structures and

those currently working on advanced theoretical models of protein structures, who are now beginning to understand why the number of possible structures that nature has 'chosen to use' is so small.

Combining Sequence and Structure (N Goldman;
D Jones, Warwick; A Lesk, Cambridge)

The final theme week in this part of the programme aimed to draw together the previous two weeks work and begin to construct a more general programme for future (*ie* beyond the Newton Institute) research. This was an attempt (albeit not the first) to draw together the approaches based solely on protein sequence and those based on analogy with known protein structures and functions. This is a difficult topic, and it was consequently not easy to make great progress. We feel that nevertheless the week was well spent, as leading researchers in related problems were drawn together and compelled to remind themselves and each other of the 'bigger picture' behind their individual research.

Workshop: Introducing Mathematicians/Statisticians to Current Problems in Biomolecular Sequence Analysis
(P Donnelly, N Goldman)

It should be clear from the remainder of the report that challenging mathematical and statistical problems abound in modern genetics, and further that the range and importance of these will grow in the "post genome" world. Nonetheless, there is a shortage internationally, and especially in the UK, of suitably qualified mathematicians/statisticians working in the area. We scheduled one structured week in which introductory lectures were aimed at those with a mathematical background - the genetics of the problem was explained simply, with a focus on the mathematical challenges. The week covered all the major areas covered by the programme, and in addition the statistical problems involved in the search for common complex diseases. One afternoon a visit to the Genome Campus at Hinxton was arranged, for further presentations and a chance to look around the laboratories.

We felt the week was helpful, and feedback from those who attended was positive. Nonetheless, the extent of the barrier presented by both the terminology, and the basic science within genetics should not be underestimated. There remains an acute shortage of trained mathematicians/statisticians/computer scientists, at virtually all levels.

Genomics: Genome Structure

(23 November to 13 December)

Entire-genome sequencing is becoming increasingly feasible. The completion of the first animal genome sequence (that of the nematode worm *Caenorhabditis elegans*) was announced in the penultimate week of the programme, and follows the sequencing of numerous bacterial and viral genomes. The human genome is expected to be completed in 2003, and numerous other animal and plant genome projects are also at advanced stages. These data are now being collected at a faster rate than they can be analysed, and dealing with this explosion of data was a recurrent theme of the entire programme. The three theme weeks between 23 November and 13 December were devoted to some of the most basic problems being generated: how do we even identify and locate organisms' genes, regulatory regions, etc, within their genome sequences? In some complex organisms, these regions of interest may comprise less than 10% of the total genome (the remainder being 'junk' DNA with no known function).

Motif and Pattern Discovery
(N Goldman; G Stormo, Colorado)

A major approach to these problems has been the identification of biologically significant motifs in functionally related sequences. A motif can be as simple as a pattern of five DNA nucleotides, possibly even incorporating some uncertainty, *eg* ATC[G|C]A, or can be a much

longer pattern of amino acids, spread over large genomic regions and incorporating multiple components separated by variable sized gaps. This week of the programme was devoted to discussion of the best ways to represent such motifs, to discover them in novel genome sequences, and to search for further examples. In addition, there was comparison of the different approaches suitable for protein motifs and nucleotide motifs.

Hidden Markov Models and Related Probabilistic Methods (N Goldman; R Durbin, Hinxton;
G Mitchison, Cambridge)

During the *Neural Networks and Machine Learning* programme at the Newton Institute (July to December 1997), a workshop was held on *Statistical Analysis of DNA and Protein Sequences*. This was immensely successful, and we were delighted to be able to devote a week of our programme to a similar theme in the 'inverse' context: a programme devoted to statistical analysis of molecular sequences hosting research on probabilistic models. Many of the same participants were able to attend (particularly since this topic is one which is exceptionally well represented in the Cambridge area), and although not formally declared a workshop, a full schedule of talks was arranged throughout the week and was attended by up to 50 people. Speakers gave introductions or updates on explicitly probabilistic modelling approaches to sequence analysis, currently one of the most successful avenues of research in a number of biological problems including gene prediction, evolutionary analysis of genomes, sequence alignment and protein structure prediction.

Gene Prediction (N Goldman; S Brunak, Lyngby;
R Durbin, Hinxton)

Simply to find genes within genomic sequences requires the integration of many different signals: promoter regions, translation start and stop context sequences, reading frame periodicities, polyadenylation signals, and, for eukaryotes, intron splicing signals, compositional contrast between exons and introns, potential differences in nucleosome positioning signals, and sequence determinants of topological domains. It is highly non-trivial to distinguish between sequences that represent true genes and those that do not, and it is clear that additional work is required both to improve detection rates and, particularly, to decrease the level of falsely predicted genes. During this theme week a number of scientists who devise and use these methods presented recent developments and discussed outstanding problems. The main concentration was on probabilistic approaches, which were felt to be the natural way to handle the complexity of the problem of incorporating information from the numerous signals which to a large extent complement each other.

Workshop: Bioinformatics, Mathematics and the Genome Project: Future Challenges (P Donnelly, W Fitch, N Goldman)

On the final afternoon of the programme we arranged a high profile and widely-advertised meeting, which hoped to:

i) Give some sense to those who had not been at the programme of its structure, achievements, and importance.

ii) Look forward, with world leaders (Blundell and Brenner) speculating on future challenges within the genome context.

iii) Bring to the attention of a wide audience the important role which mathematics and statistics have to play in the field.

The meeting was oversubscribed. Even with a video link to the upper floor of the Institute, some of those who wished to attend, but applied late, were unable to do so. A concerted (and successful) attempt was made to attract senior individuals from funding bodies and interested commercial organisations, in addition to mainstream scientists. The meeting seemed successful, both in its basic aim of presenting exciting scientific developments, and in the

wider aim of making a wider community aware of the (increasing) importance of mathematics and statistics.

Achievements

It was constantly in the forefront of the organisers' minds that this programme would be of greatest benefit to both UK and other scientists if it were able to bring together molecular biologists and data analysis theoreticians. Bioinformatics and genome-based research are young sciences, and are in need of improved links between the biological problems that must be answered and the data available towards this end on the one hand, and the mathematical, statistical and computational skills which generate the answers from the data (or know the limits of what can be inferred from different data sources) and are themselves often inspired by biological problems on the other. We are delighted, on reading participants' reports on their visits to the Institute, by the very high proportion who specifically mention the inspiration they received from participants with specialities different from their own, and the new cross-disciplinary collaborations that were initiated. Inevitably, as new problems and approaches tended to be the focus of attention, a lot of the achievements of the programme are currently intangible. The success of this programme will be best measured in two to five years time, when the significance of new projects begins to be judged by the wider scientific community. Given the high hopes and expectations of our participants, themselves experienced scientists, we remain confident that the long-term impact of this programme on the field will be very high. The participating scientists' readiness to collaborate with specialists in very different areas to their own, and the Newton Institute's unequalled ability to facilitate such interactions, consistently exceeded the organisers' expectations throughout the duration of the programme.

Richard Goldstein (Michigan) was elected as Rosenbaum fellow for the duration of the programme. He was able to contribute expertise on a number of topics relating to protein biochemistry and structure, and made the maximum possible use of the opportunity to interact with statisticians and learn about recent advances in computational statistics. An example of the research he worked on during his time at the Institute is modelling the evolutionary process in proteins and the application of this to the analysis of viral phylogeny.

A related project, initiated with Nick Goldman during the Viral Evolution workshop within the programme, was a study which is pursuing an explanation of the surprisingly low effective population size (the number of viruses in a population actually contributing offspring to subsequent generations) of HIV in humans. They believe that this might be simply explained by the recently described finding that HIV is highly compartmentalised within an infected human. These two findings were each the subject of presentations during the Viral Evolution workshop, but had not previously been linked.

A fundamental component of modern bioinformatics is the scanning of exceedingly large sequence databases for potential matches to a newly-determined 'query' sequence. The score assigned to every potential 'query-target' match can be calculated relatively easily, but the statistical assessment of such a very large number of scores, each one itself optimised over the many possible ways of aligning a pair of sequences, is non-trivial. David Siegmund (Stanford) and Richard Mott (SmithKline Beecham, Harlow) fortuitously discovered that they have been working on an identical problem relating to the estimation of significance levels for these statistical tests. As well as an exchange of their own ideas, they received further useful input from Chip Lawrence (Albany) and Gary Stormo (Colorado) who are interested in related problems.

Grainne McGuire (Reading) has developed her past work on the detection of recombination in phylogenetic data sets. Recombination is the 'horizontal' transfer of genetic information

between contemporary individuals in a population, instead of the usual 'vertical' transmission by evolutionary descent. Recombination is not allowed for in most evolutionary analyses, and so can severely compromise evolutionary inferences in some cases. The first step towards ameliorating this situation is to identify sequence regions that have been subject to recombination. Dr McGuire's approach to this problem has been via hidden Markov models (HMMs), which facilitate computations but are, in biological terms, unnecessarily constrained. During her visit to the programme, she was able to discuss with David Balding (Reading), Bob Mau (Madison) and Graeme Mitchison (Cambridge) the relaxation of the HMM formulation by the use of Markov chain Monte Carlo methods. In collaboration with Nick Goldman and Ziheng Yang (London), considerable progress was made on clarifying the analysis of sequence alignments which include 'indels' (insertions or deletions of sequence regions in some of the sequences studied) and/or missing or erroneous data (due to incomplete or inaccurate laboratory work). A simple theoretical approach has been developed, and lacks only experimental data on sequence accuracy before it can be tested. They are optimistic that these data, which should be increasingly available as genome projects reach completion, can be supplied over the next months by another programme visitor, Chris Burge (MIT).

It is widely acknowledged that there is currently no satisfactory method for combining and reconciling estimates of phylogenies from different analyses containing overlapping (but possibly not identical) sets of organisms and also potentially containing disagreeing estimates of relationships. Mike Charleston (Oxford), Andy Purvis (Silwood Park) and Mike Steel (Christchurch, New Zealand) made progress with this fundamental problem, clarifying which approaches are or are not likely to achieve all the desirable properties of a successful method. They are optimistic that new methods based on Dr Charleston's 'median network' approach will meet all of their requirements.

Tom Kurtz (Madison), Magnus Nordborg (Lund) and Gesine Reinert (Cambridge) made considerable progress with the problem of incorporating realistic levels of natural selection into coalescent population genetic models. Existing work for the cases of no, or very weak, selection can not be extended to the case of stronger selection. Their new work, including ideas on distributions of genealogies conditioned on various quantities associated with the model, have now put the strong-selection approximation on a firm theoretical footing as a limiting case.

A recurring theme in the programme was the realisation that the results of novel data analysis techniques are increasingly difficult to assess objectively. Computational methods are increasingly complex, due to both advances in statistical methodology and computer hardware, and the data sets to which they are applied are increasing in both size and diversity, due to entire-genome sequencing projects and simply the acceleration of sequence determination methods. The more complicated new methods of data analysis are virtually untestable, as they address new problems with untested algorithms, novel data sets, and using computational resources not available to journals' referees etc. In certain fields, efforts are already being made to monitor these problems. For example, in protein structure prediction there is a biennial 'challenge', the CASP competition, in which novel protein structures are offered for prediction experiments before the true results are publicly released. The 1998 CASP competition was the subject of much discussion during the Protein Structure theme weeks. On two independent occasions during the programme multi-centre collaborative projects were initiated to try to regulate such problems in other areas of research. The first came about during the theme week on Population Genetics (II), and was proposed by Simon Tavare (Los Angeles) and Peter Donnelly. While computationally intensive statistical methods offer great potential for the analysis of molecular genetics data, there are serious issues in the validation of particular implementations. Agreement was reached on the need

for a centralised web resource of data sets of various types and the "correct" answers, to relevant inference questions (likelihood surfaces or marginal posterior distributions), and on the types of data which were required. Such a collection will be established in Oxford. Work on this is in its early stages. The other was proposed during the theme week on Gene Prediction, and aims to provide standard 'test cases' for algorithms designed to scan large amounts of genomic sequence to determine the location of genes, regulatory regions, etc.

Acknowledgements

The programme benefited from grants from the National Science Foundation of the USA (**http://www.nsf.gov**), the European Community, the Wellcome Trust (**http://www.wellcome.ac.uk**) and NATO, and from sponsorship from Compugen (Israel) (**http://www.compugen.co.il**) and SmithKline Beecham Pharmaceuticals (Bioinformatics Research Group, Harlow, England) (**http://www.skb.com**). We are grateful to all of these for their generosity.
We should like to thank the staff of the Newton Institute for their support throughout this six-month programme.

Back to Top