



Department of Trade and Industry



UK Bioinformatics: Current Landscapes and Future Horizons

Prepared by: ESRC Centre for Research on Innovation and Competition
Dr Mark Harvey and Dr Andrew McMeekin

Commissioned by: DTI Biotechnology Directorate

March, 2002

Contents

EXECUTIVE SUMMARY	3
Introduction: What is bioinformatics?	10
Section 1. Emergent bioinformatic science and technologies:	
locating UK capabilities	13
1.1 The generation biological data amenable to IT analysis.....	14
1.2. Databases as infrastructural resources <i>and</i> developing science.....	16
1.3 The Development of new mathematical methods of analysis.	19
1.4 ‘Low lying fruit’	21
1.5 Conclusion.	22
Section 2 Reshaping Institutional Landscapes.....	23
2.1 Key Players in Bioinformatics	23
2.1.1.Pharmaceutical TNCs.	23
2.1.2. Agri-food TNCs.....	24
2.1.3. Computing TNCs	24
2.1.4.Dedicated Biotechnology Firms	25
2.1.5. Dedicated Bioinformatic firms	25
2.1.6. Bioinformatic Tool Providers	25
2.1.7. Public Science Institutes	26
2.1.8 Non Governmental Organisations.....	27
2.2 Bioinformatics Networks	28
2.2.1. Networks for scale and capability integration	28
2.2.2. Public and private knowledge (competitive vs. precompetitive contexts)	30
2.2.3. Industrial ‘ecology’ and interfirm agreements.....	31
2.2.4. Geographical distribution of bioinformatics activity.....	33
2.2.5. Sectoral divergence or convergence	34
2.2.6. Networks for standards and interoperability.....	35
2.3 Conclusion	36
Section 3. Resources for Bioinformatics	38
3.1 Support for UK Infrastructure and Research	38
3.2 European Support for Public Bioinformatics.....	40
3.3 International Perspective: Europe vs. United States	41
3.4 The Funding of Biological Databases – An International Problem.....	41
3.5 Private Funding for Public Knowledge.....	43
3.6 Capital Markets for DBFs, DBIFs and Bioinformatic Tool Providers.	44
3.7 Government Support for DBFs, DBIFs and Bioinformatic Tool Providers	44
3.8 Conclusion	45
Section 4. Redisciplining Skills.	46
4.1 New skills for new knowledge forms.	46
4.2 Current provision.	48
4.3 Conclusion.	50
Section 5. Creating economies of knowledge: IPR, competitive advantage, and public knowledge.....	51
5.1 Flows of knowledge and economic spaces	51
5.2 Data protection.....	54
5.3 Conclusion	54

Section 6: Future Horizons	56
6.1 The Future Horizons of UK Bioinformatics –	
The ‘Visioning’ Workshop Process	56
6.2 The Workshop Consensus Scenario.....	62
6.3 Prioritising DTI Support	64
Appendix A: Questions for Scenario Generation	68
Appendix B: The Workshop Consensus Scenario	70
Appendix C: Voting for DTI Support.....	71
Appendix D. Bioinformatics Workshop Participants	73
References.....	74

EXECUTIVE SUMMARY

1. This report describes a project commissioned by the DTI and carried out by the ESRC Centre for Research on Innovation and Competition (CRIC) to consider possible future directions for the development of UK bioinformatic capability and the most appropriate modes of DTI support in this area. The project was centred on a 'visioning workshop', where participants, representing a diverse cross section of UK bioinformatics expertise, were called on to consider these issues.

The report is presented in two main parts. Current Landscapes develops an analysis of UK bioinformatics, in a global context, based on interviews with key informants¹ (most of whom were also participants at the workshop) and a survey of the available public domain literature and information provided on websites.

The analysis is developed by exploring the activities within, and connections between, five interrelated dimensions:

- developments in science and technology
- institutional context, including public science institutions, NGOs, transnational corporations, SMEs (dedicated biotechnology and/or bioinformatic firms), and Research Councils
- resource flows and funding
- skills requirements and restructuring
- economies of knowledge regarding interactions within and between public and private spheres and associated issues of intellectual property rights

The current landscape report and three 'sample scenarios' developed by CRIC were used as background input to the workshop. During the workshop itself, participants were asked to generate a 'consensus scenario' representing a desirable and credible portrait of UK bioinformatics for 2006. The five dimensions described above were used as a structuring framework to guide this process. This 'consensus scenario' and a further set of prompt questions then formed the basis for considering ways that the DTI might most usefully support the development of UK bioinformatics capability. The outcomes of this process are described in the Future Horizons section of this document.

2. Current Landscapes

In the current landscape a number of key features emerge. UK bioinformatics has a strong presence on the global stage within a number of key areas: proteomics, crystallography, and sequence databases. Public science

¹ Many thanks to all who gave their time for telephone interviews and help in improving the draft report.

institutions in the UK, both national and European, provide a significant platform, notably centred around the Hinxton campus, but supported by the emergence of centres of excellence that are currently being consolidated. There is a strong UK presence of major pharmaceutical companies, and key global players in agri-food genomics are located around the John Innes Centre. The UK also has stimulated the growth of SMEs, especially around these two main cluster magnets. The sustained growth of these capability resources is critically dependent upon long term strategic national funding resources, complementing funding at the European level. The changing nature of biology and the disciplinary challenge presented by bioinformatics has primarily been addressed by increasing encouragement of interdisciplinarity in a variety of ways, prior to the longer term requirement for a reorientation of biology towards different experimental methods and mathematical, statistical, and computer modelling.

2.1 Developments in science and biotechnology

Bioinformatics is the application and development of computing and mathematics to the management, analysis and understanding of data to solve biological questions (with links to medical, chemo-, neuro-, etc. informatics).

In describing bioinformatics as an evolving field, which involves many different types of data and analysis, there are necessarily both processes of diversification and integration taking place at the same time. Different exigencies in private and public spheres are clearly also playing a significant role in this double process. Drug discovery and diagnostics provides a driver for more rapid but narrowly based integration across bioinformatic data domains, than some public science orientation towards general explanatory models. There is nonetheless strong complementarity and mutual gains between approaches.

Of central importance is the further development of a co-ordinated and interoperable suite of diverse databases, covering genomic through to metabolomic and organism-scale data. This should be seen as both a resource/infrastructural development and as a research/scientific understanding activity: databases within bioinformatics are epistemologically multi-functional. Secondly, there are both many types of data generation – and these will no doubt change in scale and quality – and many types of data analysis, both from within biology and closely allied disciplines, and from other disciplines.

These developments are already placing significant new demands on compute power and this is widely seen as an important driver for the next generation of supercomputers. Although significant bioinformatic activity will be independent of GRID technology, a dedicated bio-GRID will become an infrastructural prerequisite for (bio-)informatics, and investment is already being made into its development.

2.2 Institutional contexts

The development of bioinformatic capabilities is embodied in a number of different types of institution. Bioinformatic networks of diverse organisations have been established through a range of alternative modes of institutional linkage.

The different types of commercial organisation involved in bioinformatics that constitute the ‘industrial division of labour’ can be characterised by their orientations to different product markets. Bioinformatic tool providers provide specialised techniques, in the form of software and / or mathematical solutions, for use in the storage, curation and analysis of biological data. The dedicated bioinformatic firms are hosts to biological databases, selling access and expertise. Dedicated biotechnology firms are involved with biotechnological product innovation. The transnational corporations (TNCs) cover the entire product innovation pipeline, and critically have the scale of operations to market and distribute products, which could be new seeds or drugs for example. Although there are firms that overlap these categories, nonetheless firms are significantly differentiated by the product markets that they are primarily oriented towards: informatic intermediary markets, markets for tangible intermediary goods, and drug or agri-food end consumer markets.

Public science institutions play a critical role in the development of UK bioinformatics. These include dedicated bioinformatics facilities, most notably the European Bioinformatics Institute at Hinxton, and distributed research activities across UK universities, supported by the MRC, BBSRC, EPSRC and Wellcome Trust.

Significantly, the development of bioinformatics has brought about new boundaries between public and private spheres, raising important questions about public and private ownership of technologies and databases.

A further question of key importance has been identified concerning the shifting centres of gravity in the geopolitics of healthcare and agri-food bioinformatics between Europe and the US.

2.3 Resources

In considering the diversity of resource flows into bioinformatics in the UK and Europe, it is clear that there are distinctive models of growth compared with the US. There is a relative scarcity of venture capital and large corporations as yet are not involved in the support of public domain bioinformatic facilities on a scale found in the US. There is a strong tradition in the UK as in Europe that public domain science is funded from public resources and non-commercial organisations. Consequently, the future growth of public domain infrastructure and science relies on a model of expanding

public revenues. If this is not the case, the UK and Europe will significantly lag behind the US in terms of bioinformatic capability. It is perhaps unwise to assume that the UK and Europe will continue to be able ‘to punch above its weight’.

In this respect there is an opportunity for further systematic exploration of alternative models for combining different resource flows. The Hinxton campus and John Innes Centre present contrasting alternatives, the former showing a sharp separation between private and public spheres with separation of resource flows, the latter an integrated combination of corporate, NGO, governmental/EU financing, complemented by a strategy for some commercialisation of public science.

2.4 Skills: Interdisciplinarity and redisciplining

The revolutionary changes in the nature of biological science and technology over the past few decades have brought about demands for a ‘redisciplining’ of biology and new forms of interdisciplinarity. The problems of skill supply are much more to do with a restructuring of disciplines than shortages in existing disciplines. Having said that, considerable changes have taken place, especially at the higher end of skill formation, and research councils have invested in this area. It has taken much longer for these changes to be reflected upstream, in schools and undergraduate courses. Here a culture change is required, if biological sciences are to become as mathematised, both theoretically and experimentally, as the physical sciences. Conversely, biological problems are presenting new challenges to mathematical disciplines, stimulating especially nonlinear modelling. The implications are both fundamental and far-reaching.

2.5 Economies of knowledge

A central issue relating to issues of ‘economies of knowledge’ and the drawing of boundaries between public, private, and hybrid spheres, - whether these be formal and regulatory or informal and practical – concerns the flows of knowledge (and people with skills) and the flow of resources. There has to be complementarity between the different economies of knowledge – none can survive on its own. Complementarity inevitably involves asymmetries in relations between public and private economies. Public and universally accessible databases are continuously being updated and then routinely transferred into large corporations where they form an asset for in-house R & D, together with in-house generated databases. It is clear that formal regulatory IPR frameworks only affect a limited if significant dimension of the knowledge and resource flows within bioinformatics. This is especially the case given that knowledge flows can occur both through information transfers and through movement of people.

Property rights are protected significantly by software engineering and technical means rather than by law in the area of bioinformatics. Bioinformatic services and products (e.g. algorithms) are moreover subject to rapid obsolescence and are tied to few fixed assets. In these markets, IPR is less effective a means of securing future income streams than strategies for continuous innovation.

In terms of data protection, the issue is far more than one of developing means of ensuring confidentiality and protecting against financial discrimination, essential though these are. If bioinformatics is going to revolutionise primary health care, both in nature and in its delivery, the key issue is the patient-carer relationship and how that can be developed to take account of these changes.

3. Future Horizons

3.1 Sample Scenarios

By way of background and to demonstrate the value of using scenarios for thinking strategically about the future, CRIC offered three sample scenarios for 2006

- Islands of Expertise: UK bioinformatic clusters prosper, but there is no Europe-wide integration or coherent strategy
- Euro-starinformatics: A fully integrated European bioinformatic capability
- Continental Drift: UK / Europe bioinformatics activity is oriented towards agri-food applications and US activities towards pharmaceuticals

These were elaborated through considering alternative paths of possible development within and between the five key dimensions of the analytical framework.

3.2 The Workshop Scenario for 2006

The objective of the workshop scenario was that it represent a desirable and credible future for UK bioinformatic capabilities in 2006 and thus provide a basis for considering the most appropriate forms of DTI initiative.

Science and technology: The vision of the future had two strong emphases. Firstly, there has been integration across the spectrum of informatics (bio-, chemo-, demo²-, enviro-), and within the 'bio-' from molecular to organism scales. Secondly, interoperability has been advanced by the establishment of quality standards, which have improved both data input and annotation. In terms of hardware, interoperability was considered primarily in terms of broad bandwidth internet based systems. In addition, different modelling and analytical techniques drawn from diverse disciplinary domains have produced significant scientific advances. Bayesian and statistical techniques have developed within bioinformatics.

² This refers to demographic data

Institutions: Institutionally, UK bioinformatic activity is pivoted around a central hub – the Hinxton campus – but there were also strong views expressed and recorded that there was a need for other interconnected centres of excellence. This point is further elaborated below in relation to funding strategies. There was also a strong view that bioinformatics is essentially international, not bounded by regional or geo-political contexts, possibly reflecting a view about the ‘universality’ of science. So, there is no strong national or regional institutional context, and some explicitly objected to a European frame of reference.

Resource flows: There has been continued strong public funding for public science institutions, this being a powerful tradition within European countries. NGOs have also maintained a high level of funding for public science activity. Some funding for pre-competitive activity has provided a model in limited areas. Resource flows generated by the private sector are retained within the private sector.

Skills: A start has been made to the long-term goal of re-disciplining biology, to take account of the need for mathematical skills and new forms of experimentation. This has involved changes in curriculum at the very earliest stages, right through to a restructuring of university departments. In the meantime, interdisciplinary exchanges and groupings have been fostered to bring to bioinformatics methods and theoretical tools from other science backgrounds.

Economies of knowledge: There is a strong separation of the public and private spheres: public domain generated bioinformatic knowledge is maintained in the public domain, with open access. Private domain and intellectual property was deemed appropriate only for the added value produced by commercial activities in the private sector. A strong division is maintained between a public science research agenda oriented towards fundamental science, and a commercially driven R & D development oriented towards drug discovery, diagnostics, and healthcare.

3.3 DTI Support for UK Bioinformatics

The workshop identified priority areas of DTI support for bioinformatics and those considered to be ‘most favoured’ by this workshop were as follows:

- Large UK Bioinformatics Facility
- Support for database maintenance / administration / curation
- Tax breaks for industry-based financial support for universities
- Two way industry – academia sabbaticals
- Support for post-graduate training

4 Next Steps

As a document which is the outcome of some preliminary research and an initial process of consultation, this report is designed to serve as an instrument for further consultation, from a wider constituency. The aim is stimulate further strategic thinking about the future of bioinformatics, and to assist in the process of policy formation and funding.

Introduction: What is bioinformatics?

At the most general level, bioinformatics could be defined as the application of information technology to the management and analysis of biological data in order to solve biological questions (Attwood and Parry-Smith, 1999). Under this definition, the scope of bioinformatics is broad, covering anything from epidemiology, the modelling of cell dynamics, to its now more common focus, the analysis of sequence data of various kinds (genomic, transcriptomic, proteomic, metabolomic). The question ‘What is bioinformatics?’, however, is more an historical than a definitional one. Definitions change as the science and technology changes. Currently, there is a variety of alternative definitions reflecting a number of diverse approaches, some more data-oriented, others stressing analytical and computational aspects³. It should be remembered that the term ‘bioinformatics’ was first coined in the 1960s, and the first course outline in bioinformatics appeared in the late 1970s, before gene sequence data became such a central focus (Rybak, 1968, 1978)⁴. Much bioinformatics research was being undertaken, and has continued, which is not focused on sequence data analysis.

As socio-economic scientists, our approach is informed by a number of interrelated questions. We ask what kind of scientific activity constitutes bioinformatics and how the nature of scientific activity is changing; what scientists are involved, how and where. We analyse the institutions which sustain or embody this activity, and the different economic resource flows, public and private, that nourish them, and to what extent. We question what processes of skill formation, and changing institutions of academic ‘disciplines’ are involved in enabling scientific change. We ask what and how bioinformatic knowledge becomes a tradable good, how knowledge flows are constituted between public and private domains, and how bioinformatic markets are created.

In an historical context, there can be little doubt that the vast expansion of data in the post-genome era has given bioinformatics both a new salience and a new disciplinary and technological context. New sources (e.g. microarrays) and new types of data (sequence, structure, function, images and time series) are presenting quantitative and qualitative challenges. In turn, bioinformatics, rather than being seen as a self-standing discipline, now engages biology in strong interactions with hitherto relatively distant strangers: systems and control theory, systems biology (Wolkenhauer, 2001a, 2001b), machine learning (e.g. inductive logic programmes, Muggleton, 1999), 3D imaging, silicon chip and inkjet technologies as well as software development.

³ The workshop held on July 2-3 2001 produced a wide variety of competing definitions.

⁴ Originally coined by Rybak in 1968, the broad definition given in 1978 was that it dealt with problems of biosystems: ‘any natural or human process is a sort of signal and it becomes an information as soon as it is understood by human beings through a complex machinery of integrated codes.’ 158.

At a profound level, the bioinformatic revolution means that biology is being mathematised in ways quite new to it, presenting a radical challenge to the formation of skills *within* biological sciences, as well as to establishing new interdisciplinarity: getting people to talk to each other who did not need so urgently to do so before. These issues suggest that assessing bioinformatic capability centrally involves questions of the development and reshaping of human skills. The question is more than one of simple 'shortages' of this or that type of scientist or technologist measured by known demand and existing channels of supply (Gavaghan, 2000; Reichardt, 1999).

In a period of rapid change, one should *expect* to find all kinds of different things going on under the rubric of bioinformatics, a fair amount of lack of integration between them, and no clear boundaries. From a social science or policy point of view, this is both exciting and challenging, exciting because there are new kinds of social actors and interactions between them and challenging because there are no easy targets for policy intervention. Let us take one critical example, the maintenance, curation, and development of biological databases. At present there are a plethora of private and public databases, but also a lot of grey areas, where new formal and informal 'rules and routes' of access are being established (Powledge, 2001; Biotechnology Strategic Forum, 1997; 1998). Even if a perfectly dichotomous world of private versus public ever existed, there are now new ways of combining private and public resource flows, which are defining the new 'economies of knowledge' in the field of bioinformatics. A possibility exists that commercially owned databases (Celera, Incyte), given superior resources, will outstrip public databases, however generously financed from taxation (EBI, EMBL 16 May 2001). Given that some databases – and that includes strategically important ones such as SWISS-PROT – combine free and commercial accessibility, however, policy decisions on supporting these need more complex assessments.

The emergence of biological databases and their location in different types of institution is only one prominent feature of a new landscape of institutions that have emerged with the growth of bioinformatics. Dedicated bioinformatic firms now constitute a new species within the family of dedicated biotechnology firms. Large pharmaceutical companies develop networks of licences, partnerships, alliances, and collaborations with a multiplicity of different organisations, commercial, semi-commercial, and public. NGOs play a significant role in major centres of bioinformatics (Sanger). As with the science and technology, so with the institutional arrangements, a period of rapid change involves new, often unstable, combinations of different actors. The separation of some major US and UK pharmaceutical actors engaged in genomics from those involved in agri-food – a matter of indifference to much bioinformatic research reliant on comparative genomics or analysis based on convergent or divergent evolution – is evidence of an institutional *loss* of integration at this point in time.

So, from a social science perspective, bioinformatics *is* the interweaving of the emergent science and technology, the recombination of skills and human capabilities, the channelling of resource flows public and private, and the multiplicity of

interacting institutions. For any policy actor, to make specific points of intervention requires an appreciation of the fluidity and complexity of a rapidly changing and diversely populated landscape – bearing in mind also the possible complementarities with other policy actors.

To address these issues this report will fall into six sections.

1. Emergent bioinformatic science and technologies
2. Reshaping institutional landscapes
3. Combining resource flows
4. Redisciplining skills
5. Creating economies of knowledge: IPR, competitive advantage, and public knowledge.
6. Future horizons

Section 1. Emergent bioinformatic science and technologies: locating UK capabilities

If we ask ourselves what kind of scientific activity constitutes bioinformatics, we would suggest that at this historical stage of its development, the activity is *intermediatory* in the sense that it faces many ways. For example, facing towards new forms of digitalised data generated from diverse sources, its objects of analysis are constantly shifting. Alternatively, computational or simulation methodologies face data the other way round, where new models from different disciplines may experimentally generate new types of data. Finally, bioinformatics interfaces with other informatic domains: chemo-, demo-, enviro-, medico-informatics.

At present, genomic and post-genomic databases have achieved a salience reflected below, to the extent that bioinformatics has almost been identified with the rapid growth in sequence data and the public recognition associated with the success of the human genome projects. But bioinformatics is much broader than biology at the molecular level, and its centre of gravity could be displaced by integration with other informatics, and/or other computational based bio-science. In reflecting current emphasis, we do not wish to pre-empt future changes in configuration of scientific activity embraced by bioinformatics. Developments in neuroinformatics, enzymology, ion channel data generation and electrophysiology, process dynamics, and microscopy image data are a few examples of bioinformatics activity which no doubt will extend and remodel the current 'shape' of bioinformatics⁵.

Bioinformatics appears intermediatory in another sense. From interviews, it appears clear that the current phase is ground-breaking and preparing for future theoretical and analytical developments. It is not that it is currently *pre-theoretical* or *pre-analytical*, so much as it is as yet to achieve theoretical integration and comprehensiveness.

In suggesting that bioinformatics is at the receiving end of the development of new, and diverse types of data generation, we are also signalling that there are different drivers behind data generation. Lead drug discovery, single nucleotide polymorphism (SNP) analysis, agri-food genomics of stress tolerance or nutrient dense foods (NDFs), all generate their own experimental paradigms and data. So below we analyse bioinformatic scientific activity in terms of its intermediatory role between data generation, theoretical modelling and physical experimentation.

But there is perhaps a third sense in which bioinformatics can be described as intermediatory. The development of bioinformatic diagnostic techniques not only provide 'empirical data' for possible scientific analysis, but also become a vehicle for delivery of primary health care based on genomic/post-genomic bioinformatic science and medicine response profiles

⁵ Interview with Dr Charlie Hodgman, UK Manager of Bioinformatic Sciences, GlaxoSmithKline.

1.1 The generation biological data amenable to IT analysis.

Many have commented on the exponential increase in genomic sequence data, contrasting with both the rate of growth of function or structure data, or the human resources for data analysis (Grindrod, Attwood and Parry-Smith, Nellis et al.). High throughput technologies, in particular the advances in microarray technologies exemplified by those developed by Affymetrix, Agilent and Corning, will only amplify this divergence (Moore, S.K.). Density of genes per chip is anticipated to increase from 250 in 1994 to 150,000 in 2004, with the development of high-density arrays. The technology race to develop data generation can be said to be outstripping the ability to analyse it, presenting a significant problem for the future of bioinformatics which has many ramifications, to do with the quantity, quality and reliability of data. The generation of data uncontrolled by experimental hypothesis testing can lead to high levels of data redundancy, which in turn presents analytical challenges arising from the process of generation itself. Yet, it seems clear that data generation will continue to gather pace, and, short of even more problematically suggesting ways of restricting its rate of growth, the challenge becomes one of obtaining the means, both methodologically and in terms of computer power, of handling data on that scale.

The Sanger Centre

The Centre now has 20 terabytes of data, and half a teraflop of computer power. Its genome sequences increase 4x per year, and its computer power 2x per year. Major computer manufacturers, IBM, Compaq, HP, Hitachi, now consider biology to be in the process of taking over from physics as the driver behind the development of the next generation of super-computers.

In 1999 Harold Varmus, then Director of the National Institutes of Health commissioned a report on the computer needs for biology and bioinformatics, and the resultant reports (Botstein, 1999) lay behind the decision to introduce a step-change in expansion of computer power for the National Institute of Biological Research which houses genome sequences in the USA. A similar critical decision is on the table if European and UK capacity is to respond to the inexorable growth in requirements for sequence data analysis⁶.

In addition to the sheer volume of sequence data, however, there are two further processes of data generation which characterise the post-genome era, and which

⁶ Interview with Dr Richard Durbin, The Sanger Centre

present distinctive challenges to the future of bioinformatics, in the UK and Europe, as elsewhere. At the molecular level alone, there has been a proliferation of data domains (genome, transcriptome, proteome, metabolome) and a proliferation of data levels – in the field of proteomics, this is typified by primary through to quaternary levels of structural and functional analysis (primary sequences, regular expressions, motifs, fingerprints, protein-protein/protein-nucleic acid interactions, protein classifications, etc.). At the supra-molecular level, both cell and organism are rapidly developing bioinformatic levels (e.g. the mouse atlas and its virtual mouse). This double proliferation has created an enormous wealth of data of different types and levels, typical of this phase of development, but with as yet relatively little integration between them.

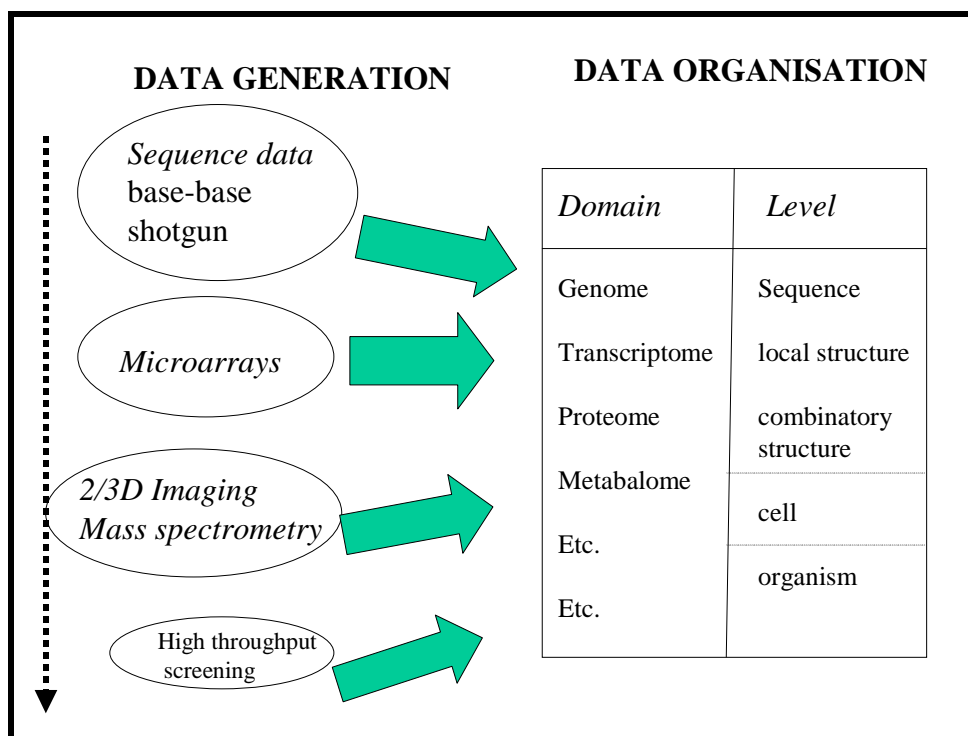


Figure 1.1 The multiple generation and diversity of data types

If one important part of bioinformatics is essentially concerned with the storage, curation, and analysis of data produced, clearly its development has been and will be profoundly affected by technologies and methods of data generation. As an *intermediatory discipline*, bioinformatics is open to all kinds of scientific and technological developments that surround it. As Attwood and Miller have observed in answering their question ‘which craft is best in bioinformatics?’, ‘biology requires a seamless integration of all the data types emerging from these fields’ (op.cit.,2), including computational biology, biochemistry, evolutionary theory, structural genomics, physiology, medical science, to name but some. These are the immediately affiliated disciplines. It is perhaps important to add that the list is far from being closed to those near neighbours, as mathematical models developed in other

disciplines such as machine learning, nuclear physics, environmental and linguistic modelling, are beginning to be reworked in the service of bioinformatics.

If Figure 1.1 represents the generation and organisation of public domain science, Figure 1.2 represents the object of this science over which the ‘seamless integration’ needs to span⁷. The challenge is for the former to be adequate to the analysis of the latter. One way of looking at the present phase is to suggest that diverse empirical probes are generating partial data representations, each to an extent locked within current respective limitations. Both the nature of the probes and the ability to coordinate them and the data they generate will eventually, it is hoped, be able to develop an analytical framework capable of spanning the object of science, however it is then conceived.

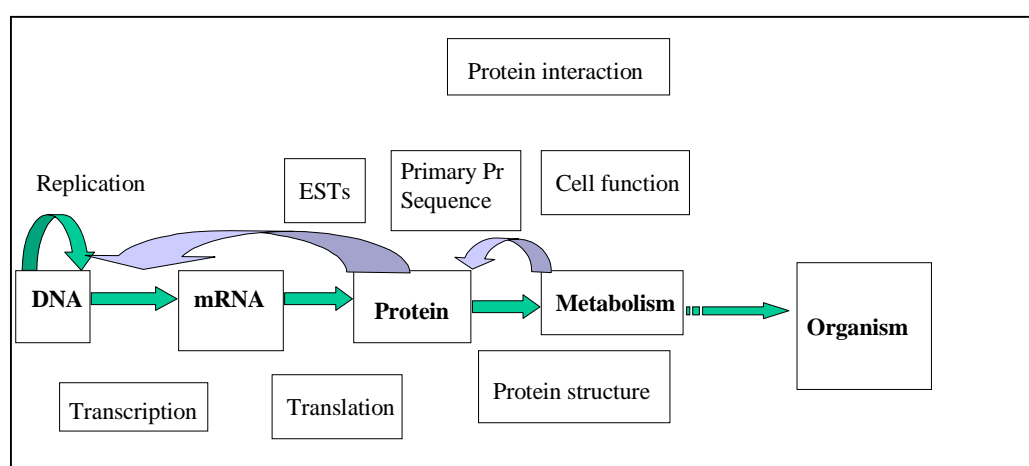


Figure 1.2. The object of bioinformatic science

If it is accepted that the current phase in the emergence of bioinformatics is one of richness in diversity, with numerous different, sometimes contending, perspectives, two main issues present challenges for the development of UK/European capabilities and orientation: the creation and curation of an organised and integrated data resource covering as full a range of data as possible, and the development of new methodologies of analysis. It is to these issues we now turn.

1.2. Databases as infrastructural resources *and* developing science

In this section, we address the scientific and technological significance of databases, dealing with the institutional aspects of where and how databases are situated and

⁷ The two figures reflect the current salience of genomic bioinformatics. As suggested above, there is no intention to bound bioinformatics within the molecular level, or restrict it to public domain activities.

funded to sections 2 and 3. It is clear that the creation and curation of genomic and proteomic databases is a core activity of bioinformatics. Moreover, if databases are not continually growing and developed, they tend to wither and die. As knowledge resources, bioinformatic databases are complex entities, performing multiple epistemological functions, empirical *and* analytical.

They are firstly repositories of the rapidly increasing volumes of data and a resource for the wider scientific and commercial R & D community, through internet access. The EBI nucleotide data base currently receives 400,000 hits a week, and provides a service for an estimated 10,000 research biologists worldwide. In order to provide this necessary infrastructural resource, standards of quality of data and consistency of annotation have to be developed. Dedicated software engineering has to be developed to make the data readily usable. Finally, interoperability between databases in order to make the data they contain comparable and combinable, has become a key issue, and one which is continuously being evolved, as discussed further below.

But secondly, the development of databases directly contributes to the advancement of scientific understanding, rather than databases simply forming reservoirs of the 'raw' data upon which analysis is based. This applies as much to the nucleotide databases, such as those resulting from the Human Genome Project, as to the proteomic databases. Thus, for ENSEMBL, developed by Sanger / EBI and located in the European Bioinformatics Institute, the human genome analysis is being extended for an annotated version of all vertebrate genomes – with particular current attention on the mouse, puffer and zebra fish. The expansion in the range of vertebrate genomic data and its annotation is itself a process of developing evolutionary and functional knowledge, which is far more than simple data accumulation. In the case of proteomic data-bases, the proliferation of levels of data mentioned above involves the development of different databases with different structural and functional information, each with their own algorithmic rules. To take two examples of secondary structure protein databases, the Pfam database at Sanger is a database of the results of applying Hidden Markov Models, whereas PRINTS is a database of 'fingerprints' of ungapped motifs produced by iterative searching of the primary protein sequence data-base, SWISS-PROT. Higher order, protein databases, such as CATH or SCOP which classify proteins according to different rules, likewise involve development of a different level of proteomic knowledge, at the molecular as against sub-molecular scale. Interviews and literature have suggested that there is a tension (see Sections 2 and 3 below) between funding for infrastructure and funding for research. Consequently database creation/curation, deemed to be an infrastructural activity, has encountered problems in attracting sufficient funding.

Given the multiple epistemological functions of genomic databases, their role and salience in UK and European bioinformatic capability is critical along several dimensions. It is clear therefore, that if the UK and Europe are to retain and develop this capability, database development across the nucleotide and proteomic ranges remains of primary importance, if the objectives of developing an integrated analysis from nucleic sequence to metabolism and organic function are to be attained. The EBI EMBL-bank database (as part of the European Molecular Biology Laboratory) is at

the core of the European capability in genomic bioinformatics, with partners in the US (Genbank at the NCBI) and Japan (DDBJ).

In the field of proteomic databases – at least in the public domain – Europe has probably gained a pre-eminent position, built around its primary protein sequence database, SWISS-PROT. Table 1.1 below lists some of the principal primary and secondary protein databases, and it demonstrates how different types of proteomic structural *knowledge* (column 1) are embedded in, and developed within, different databases from a common primary sequence data base.

Structural data type	Secondary database	Intermediary database	Primary Source
Regular expressions	PROSITE	--	SWISS-PROT
Weighted matrices	Profiles	--	SWISS-PROT
Aligned motifs (fingerprints)	PRINTS	OWLS	SWISS-PROT
Hidden Markov Models	Pfam	--	SWISS-PROT
Aligned motifs	BLOCKS	PROSITE/PRINTS	SWISS-PROT
Fuzzy regular expressions	IDENTIFY	BLOCKS/PRINTS	SWISS-PROT

Table 1.1 Types of knowledge embedded in databases (Adapted from Attwood and Parry-Smith, 46.)

But, as suggested above, we are dealing with a rapidly moving target. In recognition that no single secondary or structural protein database provides a comprehensive resource, analysis of structure and function of proteins entails increasing abilities to combine the different structural resources contained within each of the databases and their algorithmic tools. Figure 1.3 represents the current state of play, resulting from the decision to fund the next stage of integration of secondary level databases within the EBI. This in turn reflects the centrality of the Sanger-EBI duo for European and UK bioinformatic capability in the field of databases development. The interoperability between different secondary databases, and common linkage to SWISS-PROT and ENBL-bank can provide a strategic platform in this area (under the project Integr8, complemented by the development of macromolecular (EMSD), and protein-protein interaction (INTACT) databases).

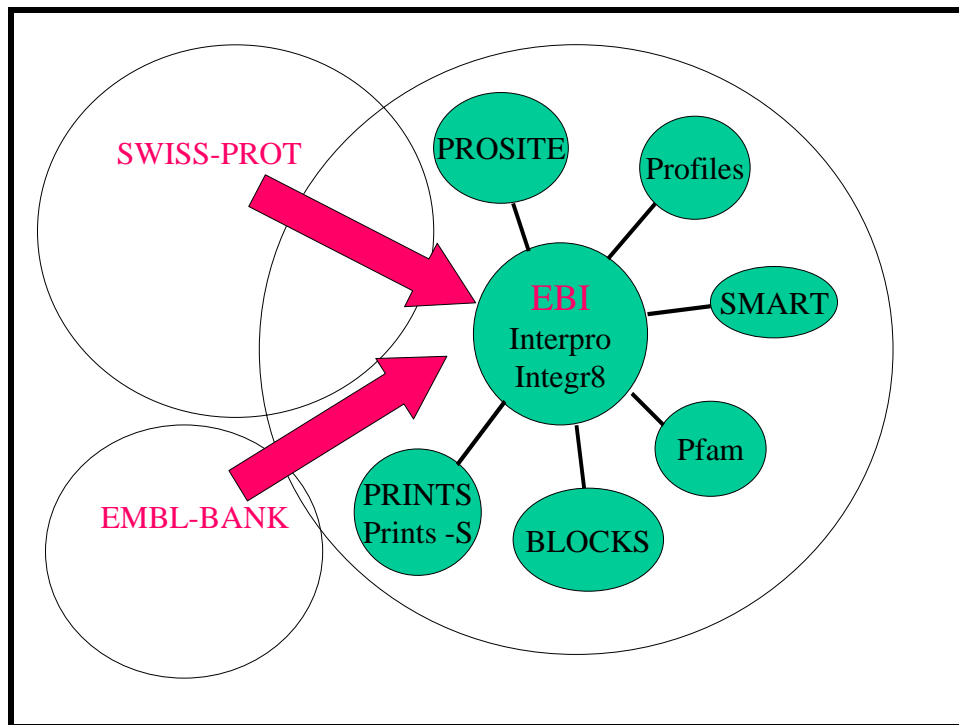


Figure 1.3. The current database European/UK platform⁸

Finally, although in its initial phase, there is the critical question of the development of GRID technology, designed to create a shared computing infrastructure (hardware, middleware, and software) with access to greatly enhanced compute power. Below we discuss the institutional aspects of GRID development. But clearly the development of interoperability between multiple and distributed databases is itself a crucial infrastructural basis of bioinformatic scientific activity. The fact that major computer manufacturing companies are now looking to the life sciences as the future drivers of development of supercomputers and high end computing suggests that the development of a European bio-GRID is of central significance, perhaps yet to be fully reflected in the e-Science programme (UKHEC, 2001). Certainly, there is a need to be cautious in assuming that a bio-GRID would simply replicate the physics-GRID.⁹ But in view of the rate of growth of bioinformatic data and data variety, a distributed computer resource and capacity is fast becoming an infrastructural prerequisite for the future of bioinformatics.

1.3 The Development of new mathematical methods of analysis.

1.3a Modelling issues related to sequence analysis. An important conclusion of Grindrod's (2001) survey of mathematical methods adopted in current genomic bioinformatic activity in the UK was that there was a bias 'towards data processing

⁸ We are grateful to Professor Teresa Attwood for providing the preliminary sketch of this diagram, as well as much of the background information in interview.

⁹ Interviews have cautioned us to reflect on the different uses and demands placed on a bio-GRID infrastructural resource, notably in terms of medical and primary care interfaces (Professor David Ingram, Dr Richard Durbin)

issues, rather than structural analysis and modelling.’ (op.cit.17). He usefully distinguishes between bioinformatics as data processing and access enablement on the one hand, and as mathematical modelling and problem solving on the other. There is continued tension within bioinformatics on the relative weight of these different types of scientific activity. Much of the dataset embedded knowledges of protein structure just described entail algorithms of pattern recognition and searching across large datasets. As suggested, there are differing views as to the importance of this kind of mathematical analysis. But there can be little doubt that this kind of pattern recognition, buttressed especially by manual annotation with all the human resources that implies, will for some time constitute a central activity in bioinformatics, as one major route to structural and functional understanding.

A central issue involved is how this building from the bottom-up through these algorithmic techniques of searching for similarities can be combined with structure prediction, or other mathematical modelling techniques, including time and space dependent ones. Structure prediction from primary sequence data has been described as the ‘holy grail’ of bioinformatics. However, as with many grails, it is not clear that this is necessarily one that will be found – even eventually. It seems that there are (at least) two ‘fault-lines’ which impede direct lines of determination from the nucleic acid sequence all the way through to metabolism and organism (Figure 1.2 above). Modes of regulation of gene networks and protein-protein and protein-nucleic acid interactions, suggest that derivation of structure and function directly from nucleic- or amino-acid sequences is only one possible model amongst others. It has been observed that a given sequence can relate to different structures and vice versa, and similar modular structures in different combinations can relate to different functions, and vice versa. In these circumstances, there are possibly different modelling routes that may be important and for this reason, the development of new computational, mathematical and statistical techniques is likely to be crucial in the long run to complement bioinformatic data-processing and management

1.3b Other new computational, mathematical and statistical techniques

Computer cell simulation, virtual organism modelling, kinetic modelling, machine learning, inductive logic programming, control systems analysis, *in numero* computational studies of gene regulatory networks, dynamic process simulation,¹⁰ are a few of the approaches being brought into bioinformatics. New types of modelling based on callibrating graphs using protein interaction and other gene interaction databases have been developed and patented¹¹. The list is not intended to be exhaustive, and clearly there are also mathematical approaches from other disciplines and domains that remain to be adopted and developed within the area of bioinformatics. Modelling techniques drawn from different disciplines (physical sciences, applied mathematics, control engineering, etc.) can help the biologist in experimental design, to decide which variables to measure and what relationships or

¹⁰ We are grateful to interviews with participants for providing us with these examples.

¹¹ Numbercraft, a mathematical consultancy firm which draws on interdisciplinary approaches, have contributed to this type of bioinformatic activity as a distinctive type of bioinformatic commercial organisation.

response pattern to look for¹². There is a widely shared view that interdisciplinary collaboration, where ‘imported’ mathematical modelling techniques will need to be *significantly modified* for dealing with biological data, will require some fundamental changes in biological assumptions on the part of biologists, and mathematical assumptions on the part of the ‘import’ disciplines. But that is a significant aspect for future developments within bioinformatics. The process of recombination of different knowledges will involve their transformation.

It should be stressed that different types of bioinformatic activity – to simplify, data- or mathematical model-driven – represent competing views of the strategic directions that may be taken. Whilst there is no reason to believe that one type of activity necessarily develops at the expense of the another, it is important to recognise that there is competition for both scientific recognition and resources between different practitioners of bioinformatics in both the academic and commercial domains. There are not only races *within* one type of bioinformatics (e.g. to complete genomes of particular species), but *between* types of bioinformatics. Indeed, this competition can be seen as a significant dynamic in the proliferation of a diversity of directions taken by bioinformatics which in the long run drives and shapes some overall advances in biological understanding.

1.4 ‘Low lying fruit’.

The discussion above has largely but not exclusively focussed on public science developments. It is clear, however, that there are also different routes and pathways, which in turn feed into the central issues described above, deriving from drug discovery and development, on the one hand, and agri-food genomics on the other. In describing these as ‘low lying fruit’ there is no disparaging implication intended that they are especially easy – or cheap – to harvest. But, a different type of data generation and analysis, often spanning the whole ‘figure 2 scientific object’, is involved in drug development by pharmaceutical companies large and small, and by the development of plant and animal species genomics for agri-food. To take one example, using single nucleotide polymorphism (SNP) databases combined with family data across Europe, data can be generated contrasting disease bearing populations from disease-free. Target genes can be identified, and expression data used as a basis for identifying protein sequences against which drugs can be tested. Such a directed pathway involves synthesising knowledge from sequence to function, in a narrow but efficiently focussed channel¹³. This too yields a different type of discriminatory and comparative analysis through a focussed channel. Lead drug discovery is an especially significant development, because of its experimental paradigm of analysing interactions between target protein functions and chemical agents, thus increasingly combining bio- and chemo-informatics.¹⁴ These pathways may be less comprehensive in terms of their yield, but nonetheless have compensating possibilities of producing integrated understanding more rapidly. The type of knowledge produced through this oriented (but also fundamental) research and

¹² Note from Dr Olaf Wolkenhauer.

¹³ Interview with Dr Chris Rawlings, Oxagen.

¹⁴ Interviews with Dr Mark Swindells, Inpharmatica, and Dr Tom Flores, Synomics.

development can also provide a significant input to the types of data and analysis which aim directly for comprehensiveness. This ‘channelled’ approach thus contributes to the diversification of routes – there are many R & D activities, many ‘channels’ being developed independently of each other.

Finally, it is clear that there is extensive mutual dependency and complementarity between this privately generated science and technology and the science and technology generated in the public sphere. The development of each depends on the other.

1.5 Conclusion.

In describing the field of bioinformatics as an evolving process, which involves many different types of data generation and analysis, there are necessarily both processes of diversification and integration taking place at the same time. Different exigencies in private and public spheres are clearly also playing a significant role in this double process. It is clear that each feeds of the other – no integration without diversification, and vice versa.

In terms of UK and European capabilities, therefore, it is clearly essential to recognise the rapidly changing nature of the ‘beast’. Of central importance is the further development of a combined and interoperable suite of diverse databases, covering genomic through to metabolomic data and beyond. This should be seen as both a resource/infrastructural development and as a research/scientific understanding activity: databases within bioinformatics are epistemologically multi-functional. Secondly, there are both many types of data generation – and these will no doubt change in scale and quality – and many types of data analysis, both from within biology and closely allied disciplines, and from other disciplines. There is no one golden route, but a necessary combination between different analytical methodologies. The aim must be to foster fruitful combinations.

Section 2 Reshaping Institutional Landscapes

Current international bioinformatics activity is developing in institutional networks comprising organisations from both the public and private sector, contributing expertise in a range of areas. Bioinformatics innovation is remarkable in its requirement for the creation of synergies between hitherto distinct capabilities. We are interested in analysing the changing character of institutions involved in bioinformatics, how the public / private divide is maintained or reconfigured and how interactions occur between different types of institution. A critical issue is to consider how the private domain depends on the public and vice versa and this clearly links forward to section 3 on resource flows. The institutional presence of bioinformatics also involves major changes in both internal institutional organisation and ‘industrial division of labour’. This section begins with a brief description of the types of institution involved in bioinformatics, with some examples (2.1), before turning to the nature of bioinformatics networks and a discussion of the main issues that impact on their activities (2.2).

2.1 Key Players in Bioinformatics

Transnational Corporations (TNCs)

There is now a sharp distinction between the activities of some TNCs active in healthcare and agri-food markets. This is perhaps indicated most clearly by the recent formation of Syngenta, the result of a merger between Zeneca Agrochemicals and Novartis, and simultaneous demergers of these divisions from their parent firms, which are now dedicated to healthcare activities. This recent bifurcation in industrial application of biotechnology platforms is widely thought to have arisen through increasingly significant differences between the markets that the firms operate in. However, it is perhaps less clear whether the impacts of these different market conditions have or could have an impact of the use of bioinformatics in R&D activities (we return to this in later sections).

2.1.1. Pharmaceutical TNCs.

Research and development expenditure by the pharmaceutical industry in the UK amounts to more than £2.5 billion a year. On average, it takes around 10 to 12 years and £350 million to develop a new medicine.¹⁵ Whilst a considerable proportion of this expenditure falls with the development phase, the discovery end is still significant. It is within this stage, that bioinformatics has its primary impact; it is hoped that drug discovery will become more efficient through the use of these new techniques, reducing the time and resources required.

Pfizer Group R&D (PGRD)¹⁶ employs 12,000 staff and had an annual expenditure of \$4.7 billion in 2000. The main European site is in the UK (at Sandwich) and employs 3,000. Bioinformatics falls within the Discovery Group with responsibilities at the

¹⁵ Association of British Pharmaceutical Industry (www.abpi.org.uk).

¹⁶ Information from www.pfizer.com and interview with Jerry Lanfear on 11/06/01

'front end' of drug discovery. The bioinformatics group at Sandwich was set up 6-7 years ago and recent developments have seen increased communication between the bioinformatics and chemoinformatics activities.

GlaxoSmithKline¹⁷ R&D expenditure was £2.5 billion in 2000. Prior to the merger, bioinformatics was scattered in a number of different divisions, and not institutionally recognised as a central driver articulating activity across a number of fields. Post merger, there are nine bioinformatics departments, with 150 scientists worldwide. In the UK, there are 50-60 in bioinformatics organisationally, but many more working with bioinformatics in some way or another in other departments. Glaxo Wellcome established its in-house International Committee on Bioinformatic Management (ICBM), but this did not lead to major organisational changes. So the merger clarified an organisational identity for bioinformatics.

2.1.2. Agri-food TNCs.

The development and analysis of crop genome databases provides an input to innovation in crop protection and crop quality to deliver enhanced yields and quality in crops.

Syngenta had a combined 1999 pro forma research and development investment of approximately \$760 million and over 5,000 R&D staff. It has major R&D outfits in the United States (La Jolla, California and Research Triangle Park, North Carolina) and Europe (Basel, Switzerland and Jealott's Hill, UK). Their involvement in bioinformatics is most clearly demonstrated through their participation in the recently completed Rice Genome Project (more details in Network 4, p35).

2.1.3. Computing TNCs

These are entering the frame as providers of dedicated computing hardware, software and internet technologies required for the burgeoning quantities of data handling, analysis and accessibility.

Sun Microsystems¹⁸ is a provider of computing hardware particularly aimed at networked computing. Under the auspices of their Discovery Informatics Programme, Sun have established an Informatics Advisory Council, constituting representatives from big pharma, tools providers and academics.

IBM, Compaq (see network 1, p29), Oracle and Hitachi (see network 4, p35) are also prominent actors increasingly oriented towards and involved in the development of bioinformatics.

¹⁷ Interview with Dr. Charlie Hodgman, GSK

¹⁸ Information from www.sun.com, interview with Susan Stephens and documentation sent by Susan Stephens to the authors.

2.1.4. Dedicated Biotechnology Firms

Dedicated biotechnology firms (DBFs) are commercial institutions whose primary business is the development of biotechnological knowledge and products. In many cases the majority of their activities lie within research and development, with revenues where they exist coming in the form of licence agreements, collaborative projects with larger firms or from the sale of biotechnology knowledge. British Biotech and Celltech are prominent UK examples. DBFs are normally users of bioinformatics products, and are likely to have dedicated informatics groups within their R&D operations. In some cases, DBFs have developed their own propriety bioinformatics systems (e.g. Cambridge Antibody Technologies and Oxford Glycosciences).

2.1.5. Dedicated Bioinformatic firms

Dedicated bioinformatic firms (DBIFs) have emerged more recently as a specialised class of DBFs whose core business is the development and commercialisation of bioinformatic products. These products are typically combinations of proprietary databases, software and techniques for data analysis. Celera and Incyte, both based in the US, are the most prominent and established examples of DBIFs. The main customers for these DBIFs are pharmaceutical and agri-food TNCs.

Incyte provide ‘an integrated platform of information technologies designed to assist pharmaceutical and biotechnology companies and academic researchers in the understanding of disease and the discovery and development of new drugs.’¹⁹

During the years ended December 31, 1999, 1998, and 1997 Incyte spent approximately \$146.8 million, \$97.2 million, and \$72.5 million, respectively, on research and development activities. This investment in research and development includes an active program to enter into relationships with other technology-driven companies and, when appropriate, acquire licenses to technologies for evaluation or use in the production and analysis process. Incyte have entered into a number of research and development relationships with companies and research institutions.

2.1.6. Bioinformatic Tool Providers

Commercial bioinformatic tool providers fall into two categories. First there are those that are involved in bioinformatics as their primary business. Nonlinear Dynamics²⁰ is one such firm, providing software products for the analysis of data generated through 1D and 2D electrophoresis gels and microarrays. It has been operating for twelve years. Their customers are both public and private institutions, including TNCs, DBFs, DBIFs and public science institutes. Synomics²¹ is a solutions software

¹⁹ Incyte Pharmaceuticals, Inc, Annual report (2000)

²⁰ Interview with Mr Will Dracup

²¹ Interview with Dr Tom Flores

provider for interoperability between disparate data sources. Its customers are large pharmaceutical companies and DBIFs, including Incyte.

The other type of tool provider is involved in bioinformatics as one aspect of their portfolio of activities. For example, Quintessa²², a mathematical consultancy firm, has been recently investigating the possibilities for extending their mathematical expertise to bioinformatics. To date, Quintessa has been a solution provider for clients involved in the nuclear, environmental and oil industries. Biology and life sciences would become a new field of activity for the firm. The involvement of this type of firm is based on a perceived analytical shortfall, based on mathematics, within the biological community.

2.1.7. Public Science Institutes

The European Bioinformatics Institute (EBI)²³, based at Hinxton, Cambridge is the flagship European public science institute in the field of bioinformatics. It is an outstation of the European Molecular Biology Laboratory (EMBL) headquartered in Heidelberg. The mission of the EBI is to ensure that the growing body of information from molecular biology and genome research is placed in the public domain and is accessible freely to all facets of the scientific community in ways that promote scientific progress. The EBI serves researchers in molecular biology, genetics, medicine and agriculture from academia, and the agricultural, biotechnology, chemical and pharmaceutical industries. It does this by building, maintaining and making available databases and information services relevant to molecular biology, as well as carrying out research in bioinformatics and computational molecular biology.

The EBI is organised under three Programmes: Service, Research and Industry. The Service Programme of the EBI focuses on building, maintaining and providing biological databases and information services to support data deposition and exploitation. Research and Development within the Service Programme investigates the latest methods in database design and interoperability with a view to providing the best possible information services. The EBI Research Programme has both pure and applied research activities at the leading edges of computational molecular biology. These activities include the study of molecular evolution, genome comparison, gene prediction, protein motifs, metabolic pathways, sequence-structure relationships, the application of parallel computing in molecular biology, the analysis of biomolecular sequences and 3D structures, new biological databases, and navigation tools for linking databases. The EBI Industry Programme was established to meet the special needs of the biotechnology, chemical and pharmaceutical industries, but still remain consistent with the public domain policy of the EBI. The programme aims to help industry adapt quickly to, and maximise benefits from, innovations in bioinformatics. The programme comprises training and education through regular workshops on leading edge topics in both biology and computing, plus the development of databases and services, with a special emphasis on the promotion and development of standards.

²² Interview with Dr David Hodgkinson

²³ The following is adapted from www.ebi.ac.uk

The other major UK based PSI is the MRC funded HGMP Resource Centre, also located at Hinxton, Cambridge²⁴. It has a stated mission²⁵:

- To provide both biological and data resources and services to the medical research community, with a special emphasis on those relevant to the Human Genome Programme.
- To facilitate genomic research by the provision of cost effective centralised collaborative and training facilities.
- To encourage users to share their data, information and resources.
- To encourage the transfer of technology from the academic to commercial/industrial applications.

The HGMP bioinformatics team has 18 staff members and 1 student supervised by a bioinformatics manager. The division provides a national on-line bioinformatics service, user support and training for external users. It provides support (but not resources) for the systems, networking and software for the HGMP-RC administration, biology services and research divisions.

The research councils, either separately or in combination have funded centres, projects and programmes in a number of different universities. For example:

- MRC Functional Genetics Unit and Human Genetics Unit
- BBSRC John Innes and Roslin and structural biology groups

The resultant picture is of a hub (the Hinxton campus) and multiple smaller centres of excellence allied to defined areas where bioinformatics plays a significant role.

2.1.8 Non Governmental Organisations

The Sanger Centre is a genome research centre set up in 1992 by the Wellcome Trust and the Medical Research Council in order to further knowledge of genomes, and in particular to play a substantial role in the sequencing and interpretation of the human genome. Sanger has been involved in large scale sequence activity with the notable achievement of producing one third of the human genome sequence, the largest single institutional contribution. Currently 50-100 of its staff, out of a total of 650, have bioinformatics as their core activity. They are engaged in extending nucleotide sequence analysis to developing annotated versions of all vertebrate genomes, as a reference for interpreting the human genome, an extension of ENSEMBL. The collaborative development of ENSEMBL with the EBI has been a significant activity alongside the development and curation of Pfam, their protein motif database.

Crucially, the Sanger Centre is situated at the Hinxton Campus, alongside the EBI and HGMP-RC. Through this combination, the Hinxton Campus is the main European hub for bioinformatics.

²⁴ Interview with Dr Diane McLaren and Dr Ian Viney for information about HGMP-RC and its relationship to EBI and Sanger.

²⁵ www.hgmp.mrc.ac.uk

The Cancer Research Campaign and Imperial Cancer Research Fund (ICRF) have also established research groups which have played an important role in the development and application of bioinformatics. For example, the ICRF have several relevant research groups including the Advanced Computation Laboratory, Biomolecular Modelling Laboratory, Computational Genome Analysis and Structural Biology Laboratory (in conjunction with Birkbeck).

2.2 Bioinformatics Networks

Much of the development of bioinformatic databases and tools takes place in inter-institutional networks. These networks span geographical boundaries and are situated across private and public sectors. The networks cross boundaries between wet and dry science and combine pure science objectives associated with generating improved understanding about the nature of organisms and highly specific initiatives related to direct application (often from the 'low lying fruit' described in section one). They are also established with a variety of different objectives. Distinctively new combinations of institutions are brought together in ways that exemplify the intermediary character of bioinformatics. There are a number of dimensions underpinning the activities of these networks:

1. Networks for scale and capability integration
2. Public and private knowledge (competitive vs. precompetitive contexts)
3. Industrial 'ecology' and interfirm agreements
4. Geographical distribution of bioinformatics activity
5. Sectoral divergence
6. Networks for standards

These are now considered in turn.

2.2.1. Networks for scale and capability integration

The sheer size of the task of continued sequence, function and structure determination means that it, in practice, it is highly improbable that a single institution would have sufficient resources to 'go it alone' entirely. State-of-the-art expertise involved in the use of bioinformatics is likely to become interlinked with institutions with radically diverse capabilities.

Inter-institutional networks often involve interactions around dedicated bioinformatic activities, both 'upstream' into the generation of biological data, and 'downstream' towards the development of products. Network 1 is an example of this, where the inter-institutional linkages in one venture span mass spectrometry, data management and analysis, and applications in the pharmaceutical industry.

Network 1: Network for experimental biology, database development and commercial application

In April 2001, GeneProt (a relatively new biotechnology company) opened a proteomics facility in Geneva²⁶. The 'factory' will have 51 mass spectrometers and advanced supercomputers to conduct protein analysis. The initiative is being backed by pharmaceutical firm, Novartis, and computer manufacturer, Compaq. It is hoped that the facility will be in a position to supply synthetically produced protein samples to pharmaceutical firms within a year. This demonstrates how bioinformatics is at the node of linkages between very diverse capabilities, spanning from wet science, through to computer manufacture and drug development.

There are numerous commercial networks, established to bring synergy across the required capabilities: in the case of network 1 the three firms bring proteomic, computational and pharmaceutical knowledge to the project. Similar networks have been established amongst public science institutions. These initiatives are often global in scale, with the Human Genome Project the most prominent. Hybrid networks, constituted by public and private institutions, also play a major role in the development of biological databases. The SNP consortium has already been described in section one (and will appear again later in this analysis). Another example is the recently established Human Proteome Organisation (HUPO). HUPO aims to be an enabling body, with the remit to raise the general profile of the Human Proteome Project, foster international collaboration and ensure that governments and the financial community are sufficiently informed about developments that they are able to take advantage of the project deliverables. It is clear that these different configurations between public and private actors have different objectives and exist in different institutional contexts. Some of these differences are explored in the following sections.

The above has described the formation of bioinformatics networks in a manner that assumes the activities to be 'big science' in nature. Whilst the 'scale' issue is clearly important, the range of bioinformatics activities can not be reduced to this characterisation. Equally important are the small-scale networks that are established to exchange specialist knowledge. In particular, this includes the next stages in bioinformatic tool development, using more sophisticated algorithms, which requires integration of capabilities between biology and mathematics. DBIFs such as Oxagen have developed particular capabilities which match in microcosm the skill profile of big pharmaceutical drug discovery in order to optimise network collaborations which

²⁶ 'World's largest leader facility in proteins field – Proteomics plant to open', Financial Times 24th April 2001.

include them²⁷. Collaborations can be as small as one-to-one interactions between academics; or in the commercial sphere, a mathematical consultant offering expertise to a larger biology-based firm.

Informal networks of knowledge sharing and capability integration are also crucial for bioinformatics. The Hinxton Campus, with the Sanger Centre and EBI, is recognised as a critical resource for providing opportunities for the international bioinformatics community to meet. There are a variety of conferences, workshops and training initiatives.

2.2.2. Public and private knowledge (competitive vs. precompetitive contexts)

There are an increasing number of biological databases becoming available, from both public and private initiatives. The biological data that are used for these databases originate from a combination of private and public sources, though there is not a one to one mapping. As it becomes increasingly important to cross reference different databases, the boundaries between public domain and commercial databases become blurred.

The development of Incyte's proprietary database, Lifeseq Gold illustrates an element of this boundary blurring, as illustrated in figure 2.1 below. The database includes data from the public domain and from Incyte's own sequencing activities. The key step is the use of 'expert bioinformatics' that add value to the raw sequence data.

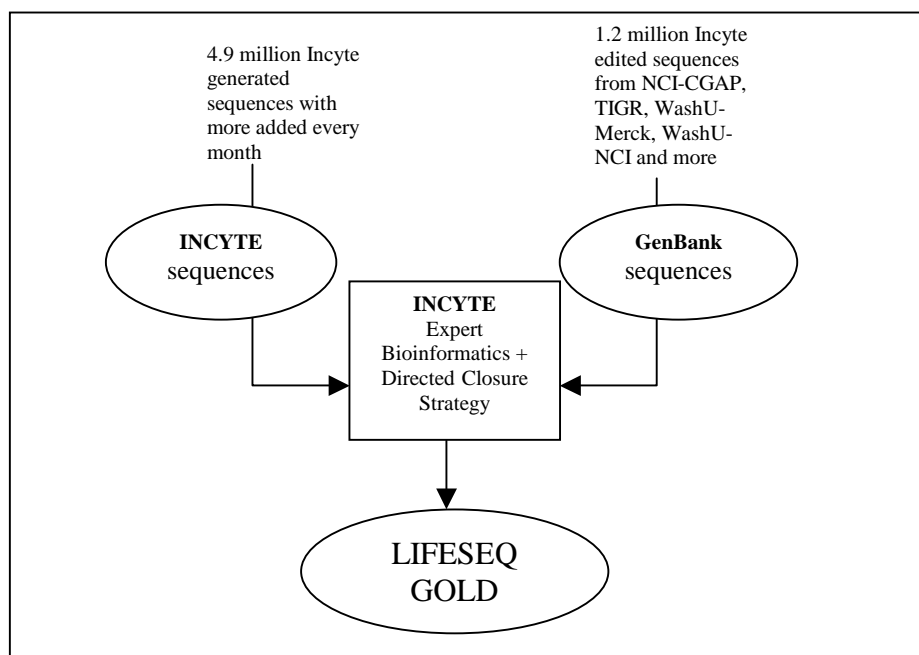


Figure 2.1: Developing a Proprietary Database

²⁷ Interview with Dr Chris Rawlings

This is one of the main components that distinguish the proprietary database from those that are publicly available (see also section 5). Indeed, if there were no added value, the Incyte business model would be distinctly vulnerable. One of the main issues regarding the private or public nature of data relates to its quality. Databases vary greatly in the extent to which they are maintained, updated, cleaned and checked for redundancy. Perhaps more importantly for future developments, databases also differ according to the amount of annotation attached to the raw data. It is possible that the private databases will need to develop considerable added value if they are to survive competition from publicly available data. In addition, with the proliferation of biological data types, there is potential for firms to stay ahead of the public domain databases. The bioinformatics groups of TNCs routinely compare the public and private data to assess differences in quality.

A further issue of public-private interaction involves the establishment of hybrid activities. Many public science institutes now rely on private income streams to complement their public funding. The result is that research projects funded privately and publicly are often undertaken in close proximity.

The boundary blurring is further magnified by attempts to create new databases through *in numero* experiments that do not use new 'wet science' data. For example primary protein sequence data might be subjected to complex mathematical modelling and analysis to produce secondary *predictive* data on protein structure or function. The use of these bioinformatic techniques can consequently add considerable value to the initial sequence data.

The question of whether certain activities are competitive or pre-competitive is closely associated to the above discussion. The traditional conceptualisation of public support for pre-competitive activities and private support for those that are commercially competitive does not appear relevant here. The formation of SNP databases has been precompetitive and as such, a global consortium of private and public institutions has undertaken the effort with considerable financial support from a charity, the Wellcome Trust. Indeed, the primary consideration for the Wellcome Trust is that all outputs from the consortium's activities be placed in the public domain. However, since Celera is also developing a proprietary database of SNPs, we must assume that there is a developing commercial component as well. The answer may lie again in the 'added value' given to the SNP data.

2.2.3. Industrial 'ecology' and interfirm agreements

The exemplar networks presented in this section indicate a range of different types of interfirm agreements. The biotechnology sector has in general seen vibrant interfirm activities. Similar activities are characteristic of the commercial bioinformatic community. There have been mergers, acquisitions, alliances, joint ventures and collaborative research activities. Another form of agreement noticeable particularly amongst the bioinformatic tool providers has been the establishment of joint marketing and distribution initiatives with TNCs. Nonlinear Dynamics, developers analytical software tools for 1D and 2D electrophoresis cells and arrays, use

Amersham Pharmacia Biotech, Hitachi Genetic Systems and Genomic Solutions to distribute their products²⁸.

An example of the shifting ecology of biotechnology companies towards bioinformatics, and the industrial restructuring that this entails, is illustrated in Network 2.

Network 2: The Changing Industrial Ecology of Oxford Glycosciences

Until recently, OGS's core business focused on small molecule and antibody drug and diagnostic products. In a shift towards bioinformatics market, exploiting their protein database, they have established a £30 million joint venture with Marconi and have taken an equity stake in US based NeoGenesis, a company specialising in high throughput screening. This complements their previous network which included Medarex, a company supplying antibody drugs.²⁹

From another perspective, Sun Microsystems, have established a partner programme, which covers activities where Sun is working with bioinformatics tools providers (e.g. Doubletwise, Timelogic)³⁰. Collaboration involves assistance in developing and optimising tools for the Sun platform. At this initial stage there is normally no major formal contract - agreements usually cover term agreements and equipment loans. Collaborations can also cover co-marketing and sales activities, where Sun can help the small tools providers by providing global coverage. At this stage formal joint marketing agreements may be established.

On a grander scale and involving large sums, DBIFs have ongoing database collaborations with a range of customers. These relationships can require the DBIF to customise the database for individual customers according to particular requirements. Network 3 illustrates that many of the large TNCs active in healthcare and agri-food sectors pay for access to this type of proprietary database.

²⁸ Interview with Mr Will Dracup

²⁹ Financial Times, 'US Group Teams up with Oxford Glycosciences', 22nd June 2001

³⁰ Interview with Ms Susan Stephens

Network 3: Networks of access to commercial databases

As of December 31, 1999, Incyte had database collaboration agreements with more than 20 companies. Each collaborator has agreed to pay annual fees to receive non-exclusive access to one or more of the databases. Some of their database agreements contain minimum annual update requirements, which if not met could result in a breach of the respective agreement. Database agreements exist with the following companies:

Abbott Laboratories	Johnson & Johnson
AstraZeneca PLC	Millennium Pharmaceuticals, Inc.
Aventis S.A.	Monsanto Company
Bristol-Myers Squibb Company	Novartis AG
Eli Lilly and Company	Pfizer Inc.
F. Hoffmann-La Roche Ltd.	Pharmacia & Upjohn, Inc.
Genentech, Inc.	Schering AG
Glaxo Wellcome plc	Schering-Plough, Ltd.

2.2.4. Geographical distribution of bioinformatics activity

Bioinformatic networks are readily visible at a variety of geographical scales, from the global, to the regional / continental, to the local. Sometimes it is assumed that simply by virtue of bioinformatics being web-based that the importance of geography has been negated. However, the physical location of databases, centres of expertise and corporate R&D facilities continues to make bioinformatics an intensely geopolitical phenomenon. As has been said many times before, location matters.

Firstly, it is true that the collaborative agreements between the major Nucleotide Sequence Databases indicate the global dimension of bioinformatics activity. The EBI, NCBI in the USA and DDBJ in Japan constitute the global deposition sites for nucleotide sequence information, and every twenty-four hours the three databases exchange information to ensure parallel comprehensivity. Mutual exchange and the carefully co-ordinated protocols required for it, contribute to the creation of the global bioinformatic scale.

But secondly, it is also true that the European Molecular Biology Network (EMBnet) represents a similar initiative at the continental level. The network was established by EMBL in 1988 to link European Laboratories involved with biocomputing and bioinformatics.

And thirdly, at the local scale, the activities around Cambridge offer insight to the importance of geographical clusters. Linked to the Hinxton Campus is a wider cluster that involves a range of key players, representing a large proportion of the types of key player detailed earlier. Technology clusters are widely believed to create a context for leading edge activities, suggesting that geographical proximity cannot be

readily substituted by communication through the internet or other means. The ability for intermittent personal contact amongst specialists often underpins the development of formal research collaborations, involving both public and private actors.

These three geographical scales clearly operate simultaneously, with particular institutions being at once local, regional and global.

The geography issue is also particularly important when considering the basis upon which TNCs make decisions to locate their bioinformatic activities, an issue of clear importance to national and regional interests. Some of the reasons relate to the location of expertise, others relate to broader institutional and regulatory issues (to be discussed in section 5). GlaxoSmithKline provide an indication of how changing circumstances can bring about shifts in location for key activities. After the GSK merger, the numbers of people engaged in bioinformatic R&D remained roughly evenly distributed between the US and Europe, but in terms of higher management a strong majority is now located in the US³¹. A similar shift has occurred within Astra Zeneca after its merger, whilst Syngenta – the agri-food corporation that emerged from corporate restructuring – retains its European centre of gravity.

2.2.5. Sectoral divergence or convergence

The earlier description of TNC activity discussed the recent split between healthcare and agri-food markets and the resultant demergers taking place amongst some of the large firms. However, bioinformatics activity occurs upstream, at several stages of removal from these end markets. The intermediate markets of the DBIFs and bioinformatic toll providers are distinct in many ways from the consumer markets of the TNCs. For this reason perhaps, it is currently not uncommon to find DBIFs and DBFs operating in both sectors and forming partnerships with TNCs in both agri-food sectors and healthcare. The implication is that the core bioinformatic competence held by DBIFs is equally and concurrently applicable in the healthcare / pharmaceutical and agri-food sectors. Network 4 shows how Myriad Genetics has been forming networks in both sectors.

³¹ Interview with Dr. Charlie Hodgman

Network 4: Networks in Healthcare and Agri-food

Myriad Genetics, Oracle and Hitachi have formed an alliance which aims to map all human proteins within three years³², with Oracle and Hitachi providing software and hardware capability respectively. The alliance is initially financed through a \$185million investment by Myriad, with the other partners and a Swiss venture capital funded together contributing the same.

Myriad Genetics, for example, were partners with Syngenta in the rice genome project, the results of which were published in February 2001³³. The project, completed in one and a half years, involved the high throughput DNA sequencing facility in combination with their bioinformatics capability:

‘The basecalling and sequence assembly software as well as software that constructs individual chromosomes from the segmented data were all developed in-house at Myriad and are available only to Myriad and its collaborators’.

2.2.6. Networks for standards and interoperability

Interoperability between databases has emerged as a critical bottleneck in the continued development of bioinformatics. Various initiatives have been set in motion to solve the incompatibility problems as discussed in section 1. The development of integrated networking and software platforms on the one hand and frameworks that allow for standard annotation, curation and querying modes for database will be critical for maximising access to the available data. It is clear that the internet in all its present manifestations, was only made possible by the establishment of globally agreed protocols and standards. In the course of that development, a number of alternative models fell by the wayside. The same will undoubtedly be true for bioinformatic databases. The two examples described in network 5 and 6 (p36) describe attempts to generate standards for database curation and software systems respectively.

³² ‘Comprehensive Map of Human Protein to be Drawn’, Financial Times, 5th March 2001 and www.myriad.com.

³³ ‘Myriad Genetics and Syngenta complete rice genome map’, Myriad Genetics press release, January 26th 2001.

Network 5: Interoperable Informatics Infrastructure Consortium (I3C)

I3C was created in January 2001 in response to wide expressions of interest from the life science community made through the Biotechnology Industry Organisation, Sun Microsystems' Informatics Advisory Council and the National Cancer Institute amongst others. The stated question and challenge posed to I3C is as follows: 'Can pharmaceutical, chemical and agricultural industries in conjunction with the academics develop such [interoperability] frameworks, in a reasonable period of time'. The primary ambition of I3C is to make available interoperable software solution sets that will:

- function within or across specified domains
- have certain agreed characteristics
- be available by the milestone dates in the I3C roadmap
- be fully interoperable independent of vendor
- satisfy the specified Domain Use Case solution sets³⁴

Network 6: Gene Ontology (GO)³⁵

The objective of GO is to provide 'controlled vocabularies for the description of the molecular function, biological process and cellular component of gene products (e.g. protein or RNA). The development of such a vocabulary is seen as a step along the path towards the unification of biological databases, by providing the means for attribution and querying at different levels of granularity.'

A core international community of academic researchers is developing GO. Other databases contribute to GO, expanding the vocabulary and refining the terms.

2.3 Conclusion

The development of bioinformatic capabilities is embodied in a number of different institutions and their interlinkages. In the process of institutionalisation of bioinformatics, new boundaries have been formed between public and private spheres, and a new industrial ecology is emerging. This new landscape raises important questions about public and private ownership of technologies and databases

The different types of commercial organisation involved in bioinformatics that constitute the 'industrial division of labour' are best characterised by their orientations to different product markets. Bioinformatics tool providers provide specialised

³⁴ Interview with and material provided by Ms. Susan Stephens

³⁵ www.geneontology.org

techniques, in the form of software and / or mathematical solutions, for use in the storage, curation and analysis of biological data. The dedicated bioinformatic firms are hosts of biological databases, selling access and expertise. Dedicated biotechnology firms are involved with biotechnological product innovation. The TNCs cover the entire product innovation pipeline, and critically have the scale of operations to market and distribute products, which could be new seeds or drugs for example. Although there are firms that overlap these categories, nonetheless they are significantly differentiated by the product markets towards which they are primarily oriented towards: informatic intermediary markets, markets for tangible intermediary goods, and drug or agri-food end consumer markets.

Finally, a question of key importance has been identified concerning the shifting centres of gravity in the geopolitics of bioinformatics. In this respect, the Hinxton complex has achieved a global standing, which needs to develop if it is to be sustained.

Section 3. Resources for Bioinformatics

Bioinformatics activity in academia and newly formed firms is funded through a variety of public and private streams. This section will consider the scale and targeting of resources from the UK research councils and charitable trusts, the European Commission, the UK Department of Trade and Industry, venture capital funds and stock markets. The targets for such support include infrastructure computing requirements, public and private databases, and public and private research (note: it is often difficult to disentangle resources that are specifically for bioinformatics and those that are for experimental biology with a bioinformatic component). Figure 3.1 summarises the primary resource flows for UK bioinformatics.

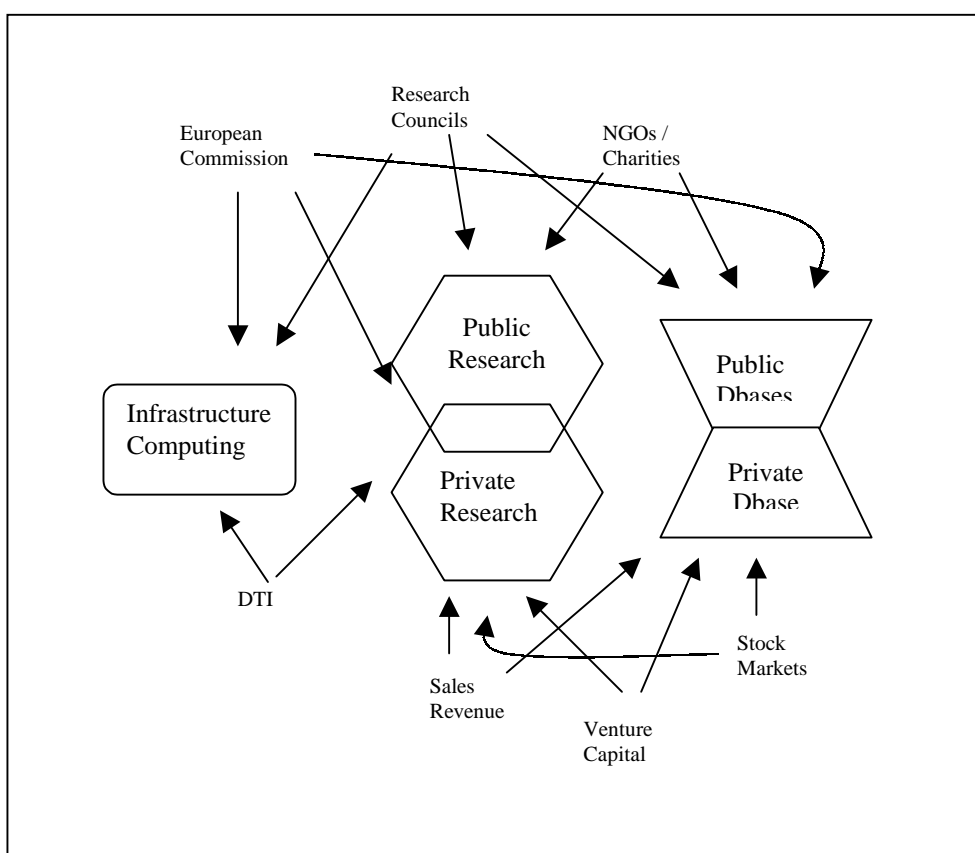


Figure 3.1 resource Flows for UK Bioinformatics

3.1 Support for UK Infrastructure and Research

In November 2000 the Director General of Research Councils, Dr John Taylor, announced £98M funding for a new UK e-Science programme³⁶. The allocations were £3M to the ESRC, £7M to the NERC, £8M each to the BBSRC and the MRC, £17M

³⁶ These details are taken from www.research-councils.ac.uk and UK HEC 2000

to EPSRC and £26M to PPARC. In addition, £5M was awarded to CLRC to 'Grid Enable' their experimental facilities and £9M was allocated towards the purchase of a new Teraflop scale HPC system. A sum of £15M was allocated to a Core e-Science Programme, a cross-Council activity to develop and broker generic technology solutions and generic middleware to enable e-Science and form the basis for new commercial e-business software. The £15M funding from the OST for the core e-Science Programme has been enhanced by an allocation of a further £20M from the CII Directorate of the DTI which will be matched by a further £15M from industry. The Core e-Science Programme will be managed by EPSRC on behalf of all the Research Councils. The e-Science initiative is to be initially funded through a combination of research council funding, DTI and industrial support.

Although these funds are targeted at a range of scientific activities, a considerable slice will be directed towards life science activities. The funds are aimed towards developing both infrastructure and bioinformatic research capabilities. The emphasis in the e-Science programme is towards Internet enabled, large scale collaborative research activity based on access to very large data collections, very large scale computing resources and high performance visualisation techniques. The 'Grid' is envisaged as a step change development of the world wide web, providing considerably greater collective computing power and enhanced networking and accessibility capability. The BBSRC, MRC and EPSRC are all in the process of allocating funds to support the development of 'Grid test beds' for bioinformatics as part of their respective e-Science allocations. In some respects, GRID development has itself played a significant role in integrating the funding activity of the three Research Councils under the auspices of the Joint Research Council.³⁷ The other target of the e-Science funding for these research councils will be towards the development of improved bioinformatic tools. All research councils have recently started to put their bioinformatic strategies into action with calls for proposals. Broadly, their objectives are similar, to support interdisciplinary initiatives to develop data management and analysis techniques for biological data. The strategies differ only as far as they are each primarily concerned to support these efforts from the perspective of their respective research communities.

Whilst the e-Science programme is the first major strategic and directed UK initiative towards bioinformatics itself, funding has been provided by the research councils for bioinformatics within other projects for some time. This previous funding for example includes the following:

- EPSRC-BBSRC joint Life Sciences Programme.
- MRC Human Genetics Unit
- BBSRC Structural Biology Group
- BBSRC CCP projects
- MRC Human Genome Mapping Project
- MRC Bioinformatic and Neuroinformatic Fellows

³⁷ Interviews with the MRC, BBSRC, and EPSRC.

Although it is difficult to identify discrete support for bioinformatics in these activities as its prominence varies considerably, it is clear that there has been considerable support in the past for embryonic bioinformatic activity as critical components within other research projects.

Bioinformatics will continue to receive support outside the core e-Science programme, though funding of projects that contain a bioinformatics component but readily fit into other research council funding activities. This being so, it will become increasingly essential that the diverse activities are co-ordinated and managed in such a way to minimise duplication and maximise the possibilities for cross-fertilisation.

The Wellcome Trust has also been a key funder of UK bioinformatics activity through their support for the Sanger Centre and the SNP consortium. The Trust is currently in the process of allocating funds to several major new projects in functional genomics, with a total investment in the region of £20 million over five years. After the success of the SNP project, the Trust envisage providing support for other 'hybrid' initiatives, on the firm basis that any outputs be placed in the public domain.

3.2 European Support for Public Bioinformatics

The EBI (as described in section 2) is Europe's flagship bioinformatic initiative. Funding for the EBI is provided largely by the Member States of EMBL. In the case of the UK, the MRC pays the UK subscription to EMBL, and thereby subscribes to the EBI. But in recent years the EBI has had some additional support from the European Commission. Other projects are supported by contributions from the pharmaceutical and biotech industry. Although its contribution to the running costs of the EBI is modest, the Wellcome Trust also provides the facilities for the EBI on its Genome Campus at Hinxton.

Concerns that European Commission funding of bioinformatics may lack sufficient resource and the required co-ordination are captured by the following statement by Peter Kind (acting director of Health Research at the European Commission):

'It is imperative for Europe to ensure its competitiveness in this field if we do not want to become a "customer" for technologies developed elsewhere and a "consumer" of products and services provided by our competitors'³⁸.

These concerns have prompted a new funding initiative to support bioinformatic activities across Europe to be spearheaded by the EBI. New funding by the EC for bioinformatics is to be co-ordinated by EBI. The contract under negotiation is worth EUR19.4 million and will be divided between four specific projects:

- i) DESPRAD – Development and Establishing of Standards and Prototype Repository for DNA-Array Data.

³⁸ EC press release, 'Genomes: knowing more, discovering faster – Boosting Europe's capability in bioinformatics', Brussels, 16th May 2001.

- ii) EMSD – European Macromolecular Structure Database.
- iii) IntAct – a public repository for protein-protein interaction data.
- iv) Integr8 – a project to build an integrated layer for the exploitation of genomic and proteomic data.

In addition to funding the EBI, the EC also funds bioinformatic activity where it is a component in research supported under other biotechnology related programmes. The main vehicles for this have been the 4th and 5th framework programmes.

The EC funded 63 structural biology projects in the Biotechnology programme of the 4th Framework Programme (1994-1998), and a pan-European co-ordination initiative was implemented in structural biology in order to improve synergy between the national research efforts. In the 5th Framework Programme (1998-2002), structural genomics is included in the “*Cell Factory*” key action, and functional genomics is part of the Generic Activities of the “*Quality of Life and Management of Living resources*” programme.

3.3 International Perspective: Europe vs. United States

The scale of European support for bioinformatics can be placed in some context by considering equivalent US expenditure of public domain activities.

The US National Institutes of Health (NIH) last year spent some EUR 300 million to support bioinformatics projects. The budget of the National Centre for Biotechnology Information (NCBI) for 2000 was EUR 38 million, as compared with EUR 20 million the year before. This amount will be significantly increased for 2001, reaching EUR 48.3 million. In the five years since 1996, the NCBI has increased its budget fourfold. In comparison, the public investment in bioinformatics in Europe hardly reached EUR 100 million in 2000, including the EBI's budget for 2000 of EUR 10 million. The year before, the EBI's budget was only EUR 7.9 million. In the five years since 1996, the EBI has only doubled its budget.

It is clear from these figures that EC resources for its flagship bioinformatics facility have been considerably less than US support for its main activities.

3.4 The Funding of Biological Databases – An International Problem

This section has already detailed some of the major resource streams for biological database. We now turn to some broader considerations regarding the establishment and maintenance of public domain databases. Indeed, given the extent of international co-operation in providing public domain databases (the EMBL-bank / GenBank / DDBJ daily exchange of data being the most prominent example) it could reasonably be argued that the issues can only be properly considered by taking a global perspective. A general indication of the nature of support for publicly available international biological databases is offered by a survey conducted in 1998,

revealing the balance of support from different funding institutions.³⁹ The breakdown is illustrated in figure 3.2.

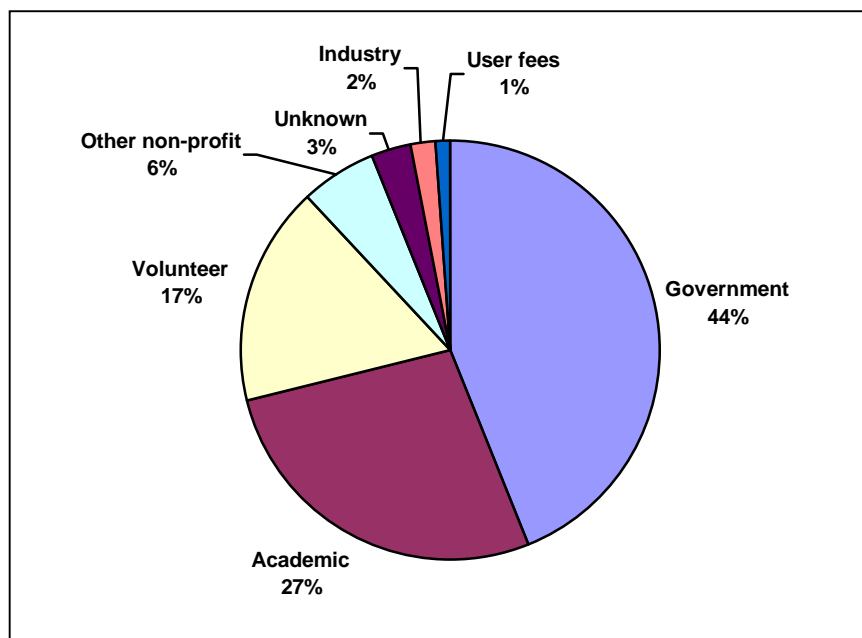


Figure 3.2 Funding Sources for International Public Domain Databases

Many of the survey respondents expressed concern about the potential for securing continuing funding, suggesting significant concerns regarding the ability of these databases to be regularly updated and enhanced. This consequently raises questions about how publicly available data should be funded in the future, throwing up a wide range of potential solutions. Furthermore, there is a current imbalance between funding the initial development of databases and funding their ongoing maintenance.

The question of who should fund public domain databases raises issues about international remit / reach and implications for the extent to which access to the data should be free or whether charges should be made.

SWISS-PROT is particularly noteworthy in this respect, as it has recently switched from providing universally free access to requiring commercial users to pay an annual subscription fee (in the region of \$90, 000 per year in 1998). For this fee, the commercial users are offered an enhanced service, SWISS-PROT Plus, which includes technical assistance, printed documentation, databases and updates on CD-ROM.

³⁹ Ellis, L. and Kalumbi, D. (1998) 'The Financial Viability of Biological Internet Resources', Nature Biotechnology. 153 databases are represented in the survey, drawn from the list of global hosts held on DBCAT.

Other possibilities that have been discussed include the income streams from advertising on websites or the use of microcommerce models where very small amounts ('pennies') are paid for each access to data.⁴⁰

These issues are further compounded by specific problems of public funding of international status databases in Europe as compared with the US or Japan. We have demonstrated in sections 1 and 2 how central the EBI is to UK and European capability. However, the politics of its funding are complex, particularly as a result of the requirement that 40% of its funding has to come from outside through short term contracts. Brought to the UK as a result of a joint bid from the MRC and Wellcome Trust, it is also dependent on funding from 16 member states within a quinquennial programme⁴¹.

3.5 Private Funding for Public Knowledge

It is now rare for public science institutions to exist without some financial input from industry. We have already seen from the discussion above that industry will contribute to the GRID initiative. Industry also contributes to research projects in academic institutes and departments. The EBI has a prominent industry programme and numerous laboratories have specific projects that are either co-funded or entirely funded through industry contributions. Major industrial actors in this respect are the large pharmaceutical companies and, to a lesser extent, the agri-food companies.

University research groups often have industry funding within their portfolios. These projects are normally characterised as pre-competitive and applied. They are usually undertaken alongside public funded basic research with standard public science objectives. It seems highly likely that there are considerable flows of knowledge between these two types of project.

The complex surrounding the John Innes Centre suggests a particularly interesting model of combining different resource flows, where substantial *corporate* investment has been made to take strategic advantage of a strong agri-food genomics cluster.

⁴⁰ Ellis, L. and Attwood, T. (2001) 'Molecular Biology Databases: Today and Tomorrow', DDT, vol. 6, No. 10.

⁴¹ Interviews with Drs Diane McLaren and Ian Viney of the MRC, Dr Richard Durbin of Sanger and Dr Alan Doyle of Wellcome Trust

Combining corporate and public resource flows

The John Innes Centre combines four different types of resource flow in order to sustain a combined cluster capability. The Centre is the focus of CropNet, the UK Plant Bioinformatics Network with Nottingham University. The Centre itself is primarily funded by the BBSRC, and is the premier European public science institute in agri-food genomics. But it also obtains NGO and governmental funding from the EU, Rockefeller Foundation, and DfID. Furthermore, it has established an important commercial wing, Plant Biosciences Ltd, and encourages a cultural mix of publicly oriented and commercially oriented research. Alongside the JIC, the Sainsbury Laboratory is funded by charitable trust, but whose commercialised outputs are also handled by Plant Biosciences. The final significant resource component comes from Syngenta (then Zeneca), who have invested £50 million in establishing a Genome Centre. An 'open-door' policy is pursued between their laboratories, the JIC and the Sainsbury Laboratory⁴².

3.6 Capital Markets for DBFs, DBIFs and Bioinformatic Tool Providers.

Start-ups and spin-offs involved with bioinformatics often rely on venture capital in the early stages of business development. Whilst it is clear that venture capital is not as readily available in the UK as in the US, it has still provided a critical source of funding for new UK firms. To date, it has been unusual for recently formed new UK firms to raise capital on stock exchanges. Several of the UK DBFs that have been operating for some time now are listed on the London Stock Exchange with notable examples including British Biotech, Celltech and Powderject Pharmaceuticals. The same has not been true for UK DBIFs. Oxford Glycosciences has a significant bioinformatics capability and provides a partial exception; it is listed on the Techmark 100 and FTSE mid 250.

3.7 Government Support for DBFs, DBIFs and Bioinformatic Tool Providers

The DTI has several initiatives for supporting new UK based firms involved in biotechnology and bioinformatics, including SMART and LINK.

⁴² Interviews with Director of the John Innes Centre, Chris Lamb and Dr Simon Bright, Zeneca, 23.6.00.

The LINK scheme⁴³ offers a framework for collaboration between the public and private sectors in support of science and technology in areas of strategic importance to the national economy. The new LINK programme in Applied Genomics, launched in July 2000, is particularly targeted at the opportunities arising in the post genomic era. It is jointly funded by the DTI, the MRC and the BBSRC, with funding of up to £15m (to be matched by industry contributions) over the life of the programme. Projects must involve collaboration between at least one science base and one industrial partner and is aimed at small firms.

The DTI SMART⁴⁴ initiative provides grants to help individuals and small and medium-sized businesses review their use of technology, access technology and research and development technologically innovative products and processes. Grants range from £2,500 to £450,000.

3.8 Conclusion

In considering the diversity of resource flows into bioinformatics in the UK and Europe, it is clear that there are distinctive models of growth compared with the US. There is a relative scarcity of venture capital and large corporations as yet are not involved in the support of public domain bioinformatic facilities on a scale found in the US. There is a strong tradition in the UK as in Europe that public domain science is funded from public resources and non-commercial organisations. Consequently, the future growth of public domain infrastructure and science relies on a model of expanding public revenues. If this is not the case, the UK and Europe will significantly lag the US in terms of bioinformatic capability. It is perhaps unwise to assume that the UK and Europe will continue to be able 'to punch above its weight'.

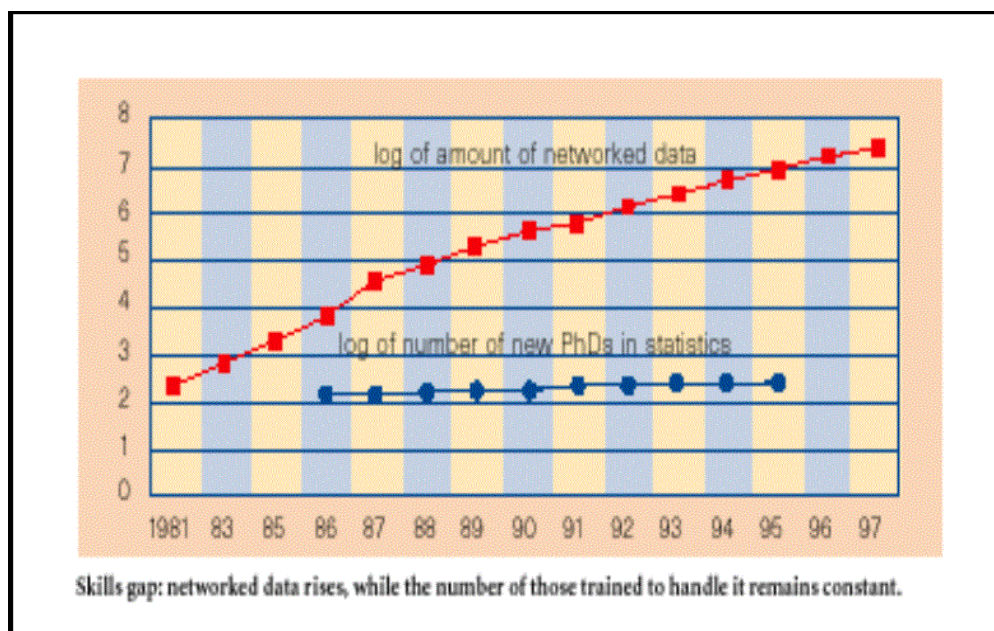
⁴³ www.dti.gov.uk

⁴⁴ <http://www.sbs.gov.uk/SMART/>

Section 4. Redisciplining Skills.

4.1 New skills for new knowledge forms.

There is a general perception, from the literature and from interviews, that there is a shortage of people with the right mix of skills. This has been presented straightforwardly as a simple shortfall of human resource outputs from education in relation to the scale of increase of data.



(Source: Reichardt, T., *Nature*, 1999, 399, 517-20)

Figure 4.1 The Skills Gap

The key issues facing bioinformatics, however, appear to be far more fundamental. A major transformation in the nature of biological science, especially the increased mathematisation of various biological areas, is changing the nature of the science. As we saw in Section 1, moreover, this change comes from both ends: an increase in data generated in a form amenable to IT treatment, and an increased availability of technologies to digitalise data (e.g. microscopy or MNR images) previously unavailable to direct IT treatment. As a consequence, the nature of experimentation and the nature of data analysis within biological science may be witnessing a similar transformation that has already occurred in physics, where experimentation begins in virtual, mathematically modelled, form as a precursor to physical experimentation⁴⁵. A transformation of the science of this order will have impacts that reach to the roots of skill formation in biological sciences.

Two immediate issues arise from the nature of this change: the nature of the skills formed within biological sciences, and of the interdisciplinarity between a transformed biology and other scientific domains. In a web survey of entry

⁴⁵ From several telephone interviews with workshop participants

requirements for physics and biology undergraduate degrees, it is clear that mathematics requirements are pitched much higher for the former than for the latter. Yet, if, as we have noted, computer manufacturers are now directing their developments to biological applications as much as to physics, ultimately that has to be reflected in education for the biological sciences. A culture change is required, to filter back up to schools, GCSE and A Level curriculae and undergraduate curriculae, to take cognisance of this change. Major pharmaceutical companies recruiting graduate biologists, such as GlaxoSmithKline, are now looking for biology graduates with bioinformatic skills, expecting a mathematical and experimental background more similar to that typically found in physics.

1.	Of 2742 BSc courses in biological science, none made mathematics a requirement.	
2.	Of 774 BSc courses in physics, all made mathematics a requirement.	
3.	Requirements for 110 BSc course in genetics:	
	Bio/Chem	30
	Bio or Chem, Math optional	19
	Science	12
	No requirements	50
4.	BSc in Cell Biology at 11 Universities, only 1 made maths an optional requirement (UCL). Main requirements biology and chemistry.	

Table 4.1 Entry requirements for degrees

(Source: www.ucas.ac.uk, June 2001)

Secondly, at least in a transitional phase, many have indicated to us that there is a need to ‘import’ skills from other disciplines, and to create new interdisciplinary groupings. Various programmes supported by research councils are addressing this issue, and in February-May 2000 there was a Joint Research Council visit to fourteen universities and subsequent report whose focus was promoting interdisciplinarity particularly between engineering and physical sciences and the life sciences (Joint Research Council, 2000). Bioinformatic aspects of life sciences figures strongly in their exemplification of the interaction between disciplines (Annex 3). The Joint Research Council has established four chairs in bioinformatics (Imperial College, UCL, Manchester, Oxford), with a focus on computational expertise. The EPSRC is also in the process of establishing long term platform research bases combining bioinformatic with physical and computer sciences. It has also been stressed that interdisciplinarity involves a two-way traffic. It is not a question of simply importing ‘ready-made’ mathematical, engineering, or physical science or statistical models from other domains. Biological functions and structures present quite distinct challenges which require novel solutions.

Conversely, biologically based bioinformaticians need to acquire knowledge and understanding of the modelling in other disciplinary areas. The MRC together with the EPSRC have recently held workshops for doctoral and MSc students in mathematics and the physical sciences to encourage career-switching by emphasising opportunities in biological sciences. Likewise, the MRC, EPSRC and BBSRC have developed opportunities for ‘discipline hopping’, for instance, supporting academics in physical or mathematical sciences to immerse themselves in bioinformatics for periods of up to a year by buying out their time. The EPSRC has also provided funds for mathematicians to attend bioinformatic theme sessions at the Isaac Newton Institute in Cambridge. It has been policy for the EPSRC to establish bioinformatic chairs in physical and computer science departments⁴⁶.

In the commercial sector, the most effective combinations of disciplines have been described as ‘hot groups’, where the different disciplines are brought together to address a common project, in a process described as ‘shepherding a flock of cats’. The ‘intermediatory’ character of bioinformatics, between data generation and physical experimentation involves a distinctive skill requirement of being ‘a jack of all trades and a master of one’.⁴⁷ There is clearly a significant contribution to skill formation and development of interdisciplinarity within enterprises, but it is beyond the scope of this report to treat it with the consideration it merits.

4.2 Current provision.

The Table 4.2 below lists the main areas of bioinformatics skill formation and concentration in UK universities in June 2001, from undergraduate through to post-doctoral level. It also includes academic posts and centres established to encourage the development of bioinformatic skills and centres of excellence. Each of the research councils funds a number of studentships and research degree courses at the masters and doctoral level, either jointly or on their own.

There are also distributed skill networks, more or less informal, for different areas of bioinformatics. For example:

- X-Ray crystallography. Centres of expertise include Janet Thornton at Birkbeck/UCL, EBI/Sanger, and Sheffield
- Inductive Logic Programming. The network includes Mike Sternberg (Imperial College to be joined shortly by Stephen Muggleton), Ross King (Aberystwyth), Oxford and Manchester.
- Database curation and development: EBI, UCL, and Manchester

⁴⁶ EPSRC website and interviews Dr Lesley Thompson, Programme Manager for the Life Science Interface Programme.

⁴⁷ Discussion and presentation ‘Graduate skills desired by the pharmaceutical industry’ April 2000, from Dr Charlie Hodgman, GSK.

There is a question of whether there needs to be an attempt to coordinate and/or recombine some of these networks.

		Institution	External Funding
Undergraduate	2 BSc Courses (1 SW) 1 BSc Course 1 BSC Course Specialist Option BSc	UMIST Birmingham Queen Mary & Westfield Abertay	Support from GSK
Masters	MSc and MRes	Liverpool Glasgow Birkbeck Kings College, London York Leeds Exeter Abertay Sheffield Hallam Manchester Nottingham Edinburgh Royal Holloway Ulster UCL	2 MRC, BBSRC 10 BBSRC, 4 MRC BBSRC/MRC/EPSRC <i>(Overall BBSRC, MRC and EPSRC each fund 20-25 Masters studentships)</i> MRC MRC
Doctoral	PhD	Glasgow Cambridge EBI Edinburgh Reading UCL UMIST	
Post-Doctoral	Research post	City of London EBI Manchester Imperial College	Marie Curie European Commission BBSRC/EPSRC
Academic posts	Chair Chair + lectureship Chair Chair Chair, reader, 2 lecturers Reader/lecturer Lecturerships 3	Imperial College UCL Manchester Oxford Royal Holloway, London Dundee Edinburgh	Joint Research Council Wellcome Trust
Centres	Bioinformatics Computing Centre Engineering/imaging	EBI/Sanger King's College London Oxford University	EU/Wellcome MRC/EPSRC

Table 4.2. Bioinformatics skills formation

(Sources: www.ucas.ac.uk and www.hgmp.mrc.ac.uk/CCP11/, June 2001)

There are also two specifically web-based centres for courses in bioinformatics, the University of Nottingham's The Virtual School of Molecular Biology, offering degrees in bioinformatics at the undergraduate and masters level, and Birkbeck College's Advanced Certificate in the Principles of Protein Structure.

4.3 Conclusion.

The revolutionary changes in the nature of biological science and technology over the past few decades have brought about demands for new types of biological skill formation and new forms of interdisciplinarity. The problems of skill supply are much more to do with a restructuring of disciplines than shortages in existing disciplines. Having said that considerable changes have taken place, especially at the higher end of skill formation, and research councils have invested in this area. It has taken much longer for these changes to be reflected upstream, in schools and undergraduate courses. Here a culture change is required, if biological sciences are to become as mathematised, both theoretically and experimentally, as the physical sciences. The implications are both fundamental and far-reaching.

Although outside the scope of this report, skill formation should not be seen as the preserve of the educational system. Not only do major companies support education and research posts within universities, but there is also a flow of personnel, the bearers of knowledge, in both directions. Furthermore, interviews revealed that flows from the public to private sector were strongly affected by salaries available in the private sector, and in turn within the private sector, salaries in the financial sector induced flows out of bioinformatics for people with informatic and mathematical skills.

From interviews, it is clear that some companies, including SMEs closely networked with universities, also play a directly educative role by contributing to courses. This kind of two-way flow might well be a fruitful area for governmental support, recognising the symbiotic gains to be made by increasing interactions for all parties. Furthermore, skill development, whether formally or informally, occurs within companies and enhances the overall capabilities within bioinformatics.

Section 5. Creating economies of knowledge: IPR, competitive advantage, and public knowledge

Bioinformatics presents some quite distinct issues related to public and private sphere knowledge, as well as to data protection, and ethical regulation of use of bioinformatic information (notably the use of bioinformatic diagnosis and individual genetic data held for medical or forensic purposes).

5.1 Flows of knowledge and economic spaces

Although there is much talk of ‘the knowledge economy’, bioinformatics illustrates the much more problematic aspects of creating distinct public or private ‘economies of knowledge’. On the one hand, it might be thought that publicly resourced science and infrastructure (including the key bioinformatic databases) would underpin public domain knowledge, so constituting a clear public ‘economy of knowledge’. On the other, protection of innovation by private sector enterprises through Intellectual Property Rights for software or algorithms might be expected to underpin a clear-cut private sector ‘economy of knowledge’. But the interrelation between resource flows (private and public) and knowledge flows (public to private or vice versa) is much more blurred and complex than that, as most dramatically demonstrated over the rough draft of the human genome developed by Celera and the Human Genome Project. The agreement between Celera and *Science* (Powledge, 2001) raised controversy over a precedent set for publication of scientific results when the data upon which it was based was not fully open to public scientific scrutiny, restrictions being placed on the total access to genomic data. This was described by Michael Ashburner of the EBI as the beginning of the “balkanisation” of genomic data.

Moreover, the situation is complicated by the divergence between EU and US IPR regulatory environments (Coleman, L. 1998, and Uhlir, P.F. 1998, Brown et al, 1999, 28-30), raising considerable controversy especially in relation to bioinformatic databases discussed in *Bits of Power* (National Research Council, 1997). Thus, the Directive 96/9/EC of the European Parliament was designed to strengthen opportunities for copyrighting databases and their development, in a way which was quite antithetic to the US legal framework, giving insufficient insurance either on data protection or on free data availability for research and educational purposes. Yet, at the same time, publicly funded databases such as SWISS-PROT, were creating a two tier system of free access to academics and commercial access for the private sector (Cameron, 1998), following copyrighting in 1998.

This latter development reflects an important aspect of different economies of knowledge, and different logics of sustainability. SWISS-PROT, as with all major databases, are extremely expensive to maintain, curate and develop. If knowledge flows go direct from public to private, the burden of cost falls one-sidedly on taxation, whereas the benefits of knowledge are gained by both private and public. Further, if, as is normally the case, private organisations then create ‘mirror’ databases which are then firm-internal assets, subsequently developed and acquiring their own dedicated analytical and software tools, a private knowledge base is created which in turn can be

used to generate commercial income from products and services derived from it. Thus, although the 'race' between Celera and the Human Genome Project for the rough draft of the human genome was presented as one between commercial and public domain, Celera's achievement also rested on importation of all relevant public domain databases. These 'mirror' databases⁴⁸ are protected by technical⁴⁹, rather than legal, barriers known as 'firewalls'. Within the firewall, knowledge is private in the double sense of being privately owned *and* not commercially available. Interoperability *behind* the firewall becomes a critical issue, to make maximum use of internally generated and 'imported' public domain databases. This has been described as akin to creating embryonic 'mirror GRIDs' within the large pharmaceutical companies, an internal private infrastructure to match the external public infrastructure⁵⁰. SKB recently has also created internal project-based spin-outs which in effect create almost a firewall-within-a-firewall system of knowledge ownership, with the spin-out firm trading as a separate entity with the parent firm (See Figure 5.1).

Tradable bioinformatic knowledge appears in the open market, and this may or may not be subject to IPR. From interviews and websites, it is clear that some software packages, together with database access might well be patentable (e.g. Incyte's LifeSeq Gold or Synomic's Lead Discovery), and is traded as such, as one business strategy. For others, proprietorial ownership is accepted as being much more transitory in tradable knowledge, and derives from competitive advantage alone, either because legal protection is too costly (e.g. Nonlinear Dynamics), or as an optimum business strategy of 'staying ahead of the game' (e.g. Oxagen).

The following diagram illustrates in a schematic manner some of the complexity and the essential asymmetry of knowledge flows between the different 'economies of knowledge.' Five economic spaces are pictured (not to scale): public domain, pooled public/private domain (e.g. the SNP consortium), open market private domain under IPR, open market private domain under competitive advantage, closed private domain behind the firewall, and closed private domain of corporate spin-outs, firewalls behind the firewall.

⁴⁸ In the SATSU survey (Brown et al), all pharmaceutical companies except one maintained mirror databases, and for many of them, public databases were their major source over proprietary databases, accounting for over 50% of their use.

⁴⁹ Through encryption or source code protection, for example.

⁵⁰ Dr Tom Flores, interview.

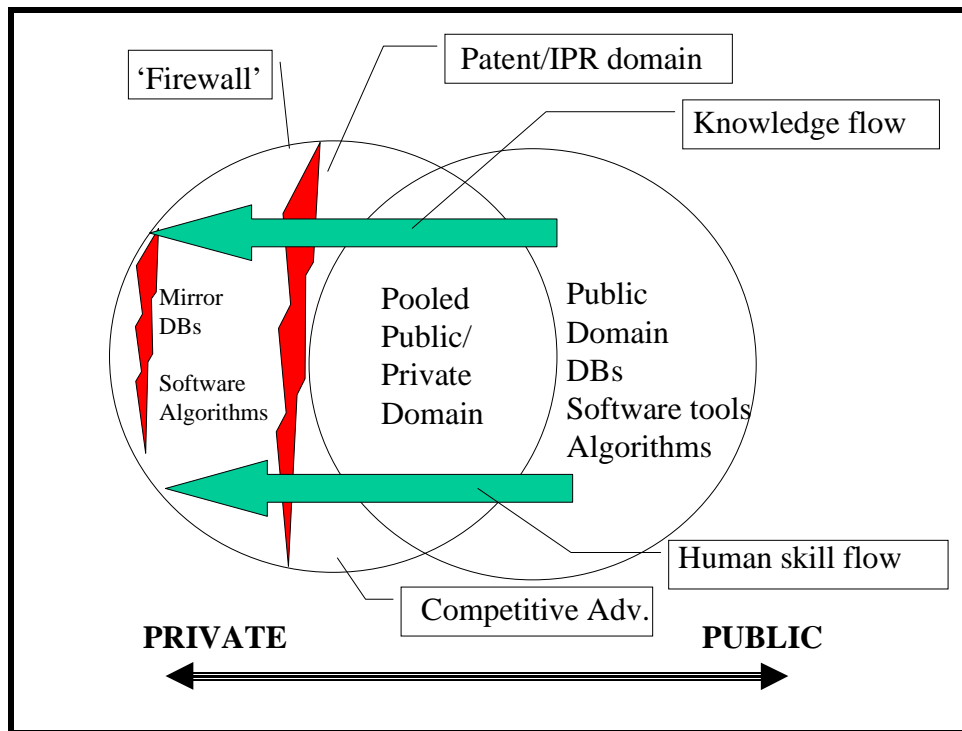


Figure 5.1. The construction of economies of knowledge between the private and the public.

It is often assumed that competition within the private sector drives firms to innovate and maintain competitive advantage. However, especially in the field of bioinformatics, it has been strongly suggested that SMEs especially are subject to strong competitive pressures coming from the public sector, which enjoys an inherent advantage of open exchange of tools and algorithm developments. As human capabilities are a key asset, it follows that the broader the capability pool, the greater the probable advantage particularly over smaller organisations dedicated to a narrow but strong focus on a particular market. The free exchange of knowledge within academia, possibly to be strengthened by GRID technology, is thus an important consideration in assessing commercial capability except with respect to the larger pharmaceutical or agri-food corporations whose budgets often exceed those of the research councils that sustain public economies of knowledge. Both IPR and competitive advantage around particular knowledge products and services are vulnerable to obsolescence. Obsolescence in knowledge markets is of a particular kind, partly because the development of a new mathematical or statistical analytical tool may not be embedded in more durable and expensive-to-replace hardware or other fixed capital. A counter-tendency to that, however, which has also been widely stressed, is that there is a continuous human skill flow from public to private sector, given the competitive advantage held by the latter in the labour market (see diagram).

5.2 Data protection.

Bioinformatics will undoubtedly have a revolutionary effect on diagnosis and therapy in the delivery of health care in coming years. This raises central regulatory and ethical questions (MRC, 2001), quite distinct from those covered by property rights issues, and concerns issues of confidentiality and rights of patients to knowledge. As in the other domain, the issue is not only one of formal rights. Without engagement of patients and/or patient organisations, bioinformatic information is likely to remain an effectively closed door. In this respect, Poortman's Centre for Patient Population Genetics in The Netherlands is exemplary, and there is a growing development of such networks through the European Patient Genetics Forum network, and the UK Genetics Interest Group.

As bioinformatics extends right down into primary care delivery, where GPs will have access to and communicate information, there is a clear need for clarity and consistency in the sensitive handling of genetic information.

The 'Icelandic case' in which the Icelandic government gave a commercial company, DeCode Genetics, the rights to establish and commercially exploit a genetic health register, together with Hoffman-LaRoche, highlights developing issues of confidentiality versus commercial/public domain use (Brown et al.)

A further key area of concern will be the use and availability of individual genetic bioinformatic systems for insurance purposes which are currently covered under the Council of Europe's Convention for the Protection of Human Rights and Dignity of the Human Being. Given the international accessibility of bioinformatic systems, it is clear that regulatory systems need to be consistent between nations, and indeed globally. In the UK, parliamentary committees have been established to monitor and advise on the development of regulation and codes of practice in this area.

5.3 Conclusion

A central issue relating to issues of 'economies of knowledge' and the drawing of boundaries between public, private, and hybrid spheres, - whether these be formal and regulatory or informal and practical - concerns the flows of knowledge (and humans with skills) and the flow of resources. There has to be complementarity between the different economies of knowledge - none can survive on its own. Complementarity inevitably involves asymmetries in relations between public and private economies. It is clear that formal regulatory IPR frameworks only affects a limited if significant dimension of the knowledge and resource flows within bioinformatics, and, of themselves, they could never achieve comprehensive regulation. This is especially the case given that knowledge flows can occur both through information transfers and through movement of people. There is therefore probably no regulatory solution to

imbalances and diseconomies, but a need to find collaborative solutions between actors within the different economies of knowledge.

In terms of data protection, the issue is far more than one of developing means of ensuring confidentiality and protecting against financial discrimination, essential those these are. If bioinformatics is going to revolutionise primary health care, both in nature and in its delivery, the key issue is the patient-carer relationship and how that can be developed to take account of these changes.

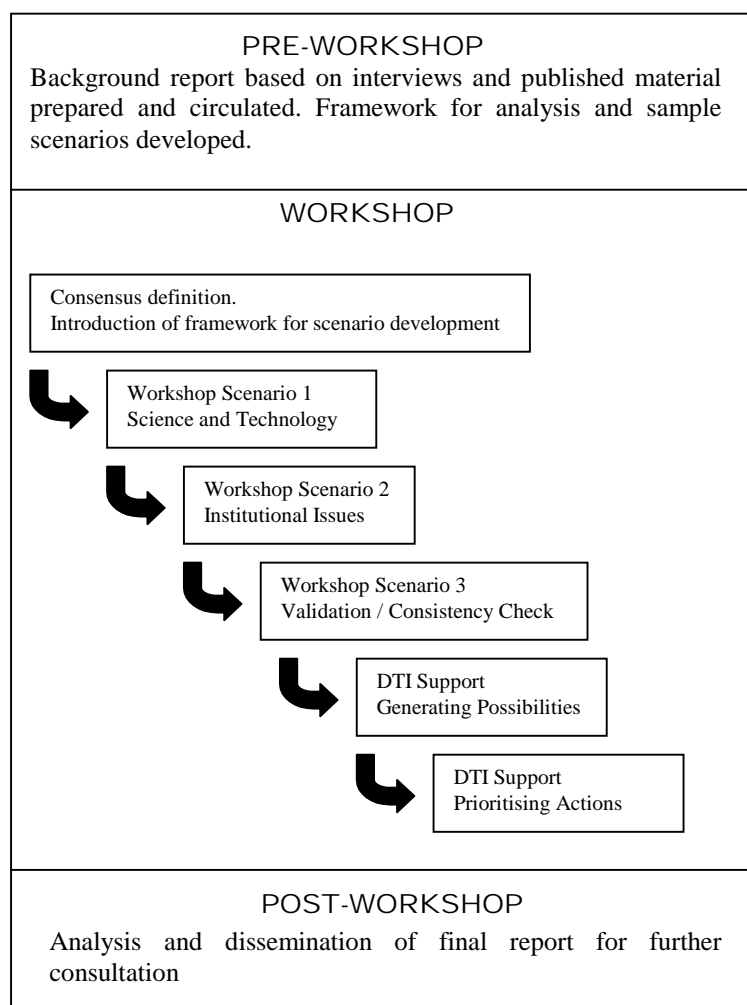
Section 6: Future Horizons

6.1 The Future Horizons of UK Bioinformatics – The ‘Visioning’ Workshop Process

The primary purpose of the workshop was to identify ways that the DTI could support UK bioinformatics generally, and specifically to pinpoint possibilities for DTI financial support. To meet this objective the workshop participants were first set the task of generating a consensus scenario, *based on the framework developed in this report*.

Scenarios are visions of the future that provide a context for strategic decision making. In particular, scenarios are a good way to expand the parameters for thinking about the future through thinking the unthinkable or the undesirable. They can be developed and used in a number of different ways, depending on the specific objectives that they are intended to meet.

In this project, the objective was to generate a consensus scenario for UK bioinformatics in 2006, in the context of global developments. In addition, the scenario was required to represent a credible and desirable portrait. This could then be used to assess the potential for the DTI to support specific initiatives aimed at facilitating progress to this outcome. The process used during the workshop is described below, as indicated schematically in the figure below.



The Framework for Scenario Development

The workshop began with a discussion of how bioinformatics should be defined for the purpose of this workshop. It was evident that there is a wide range of alternative definitions currently attributed to bioinformatics. The resultant ‘consensus definition’, for use at this workshop was intended to be as inclusive as possible:

The application and development of computing and mathematics to the management, analysis and understanding of data to solve biological questions (with links to medical, chemical, neuro, etc.)

The important differences that remain within this definition have been recounted in more detail earlier in the report.

The workshop scenario was developed on the basis of the framework used to structure the ‘Current Landscapes’ part of this report. There were five dimensions to the framework: science and technology; institutional organisation; resources; economies

of knowledge; and skill formation. To introduce workshop participants to the use of scenarios, CRIC presented three sample scenarios, again using the same five-dimension framework. These sample scenarios were developed to offer deliberately and strongly contrasting alternative visions of future UK bioinformatics. The detail, in each case, was provided by considering different possibilities within each of the framework sub-components. Crucially, the scenarios also demonstrate the importance of thinking about the links between the different dimensions.

These were:

A. Islands of Excellence: UK bioinformatic clusters prosper, but there is no Europe-wide integration or coherent strategy

Science and technology:

The UK hosts a world class vertebrate database. There have been promising developments towards interoperability *within* UK public science but poor interoperability between UK and US systems. The UK is considered to be the world leader in proteomics, x-ray crystallography and artificial intelligence.

Institutional Organisation:

Public science activities at Hinxton and the connected Cambridge cluster, including a prominent UK DBIF act as a magnet for international bioinformatic activity. There are similar, but unconnected hubs across Europe. There are pockets of excellence in university departments across the UK with globally recognised expertise. Transnational pharma companies shift the critical mass of their bioinformatics activities to the US, but retain some UK presence for connections to specific areas of UK expertise.

Resources

There is no EC strategic, large facility funding to match US NCBI, but there is strong UK funding for a central UK facility allocated through the comprehensive spending review. Patchy funding for disconnected research activities across UK is based on significant but separate NGO and Research Council initiatives.

Skills

There is a brain trickle from UK public bioinformatics to the US, and to the private sector. In addition, there is a general international trickle from the commercial life sciences to the financial services sector - largely based on salary and the availability of facilities.

Economies of Knowledge:

Public-private interaction is underdeveloped in the UK, leading to an intensification of private sphere asymmetric advantage; knowledge flows are in one direction leading to concerns over the ownership of knowledge.

B. Euro-starinformatics: A fully integrated European bioinformatic capability

Science and technology

The GRID is developing successfully at the European level providing the required infrastructure for integrated informatics, including bio-, chemo-, demo-, enviro- and medico- informatics. The UK is pre-eminent in proteomics and is also central to European-wide developments of new informatic domains (including enzymology, 3D crystallography and virtual organisms). Bioinformatic solutions from UK providers having significant impact on drug and agri-food innovation pipelines.

Institutions:

The EBI is the bioGRID coordinator accompanied by European collaboration in development of databases. The Cambridge cluster of public and private science grows and connected to other strong European clusters, including others in the UK. Pharma and agri-food TNCs are involved in collaborative research with public science institutions. A UK DBIF is a recognised global leader in proteomics.

Resources

Long term EC funding for bioinformatics research and infrastructure is secure. TNCs provide significant investment for public science - strengthening the viability of public domain databases. UK government initiatives and venture capital resources underpin an enterprise economy for bioinformatics.

Skills

There is an intensification of interdisciplinarity and redisciplining of biology. On the one hand this has involved bringing in and adapting mathematical and modeling techniques from other disciplines. On the other hand, there have been culture change in the ways that biology is taught at all educational levels.

Economies of Knowledge

There has been balanced growth for competitive, pre-competitive and public domain bioinformatics, based on creating European wide protocols for IPR, data protection, and competition policy and incentivising collaboration.

C. Continental Drift: UK / Europe bioinformatics activity is oriented towards agri-food applications and US activities towards pharmaceuticals

Science and Technology

UK and Europe has developed pre-eminence in protein - nutrient interaction. The UK hosts a world class genomic – medical – epidemiological, GRID enabled, database. The UK is world leader in Nutrient Dense Food innovation and ‘pharma foods’ are now seen as a potentially significant contributor to the delivery of healthcare.

Institutions

Drug discovery is predominantly US based on use of state-of-the-art bioinformatics with later drug development located in the UK based on imported tools. UK DBIFs specialise in agri-food bioinformatics. Agri-food TNCs collaborate with UK public

science including the establishment of interlocking private - public hybrid institutions. Supermarkets collaborate with TNCs to prepare European markets for NDFs

Resource Flows

UK public funding is stable and very targeted on integration of post genomic and medical data. European Venture capital switches to agri-food bioinformatic activities. Resource flows from European agri-food TNCs into public science across Europe

Skills

There are strong initiatives aimed at developing interdisciplinarity between bioinformatics, epidemiology, physiology and nutrition science. There is a brain exchange, as those with a pharmaceutical orientation migrate to the US and those with agri-food orientation migrate to Europe.

Economies of Knowledge

US patents on bioinformatic tools succeed, and there is no patent protection in Europe for software / mathematical solutions. New generation drugs in Europe met with regulatory and cost hurdles. Public concerns regarding the storage of human genomic data are alleviated by new European ethics code for medical data collection and protection.

In addition to introducing the use of scenarios for thinking about the future, these examples were presented to draw attention to two crucial elements of this process for the purpose of this workshop. First, that in thinking about developments in each of the five scenario dimensions, attention should be focused on how the elements are interconnected to one another so that the complete scenario is consistent. In particular, care should be taken to ensure that developments in one section are not contradictory to developments in another. Secondly, that the scenario should consider the development of UK bioinformatics, in the context of public and private activities within an international perspective.

Generating the Workshop Scenario

Two sessions were dedicated to generating a workshop scenario for UK bioinformatics, to present a desirable and credible portrait of developments by 2006. The first session involved discussions about developments in science and technology, and the second session focused on the institutional issues.

Discussions were held in small groups, each comprising a combination, as far as possible, of representatives from public science, TNCs, SMEs and research councils.

In the first session all groups were asked to consider the developments in science and technology that would constitute the workshop scenario. They were provided with a list of specific questions to answer, as detailed in Appendix A.

In the second session two groups consider developments in skill formation, institutional organisation, resources and economies of knowledge. The questions they were asked to consider are detailed in Appendix A.

Validation of Workshop Scenario

This session was dedicated to the task of considering the aggregate / consensus scenario developed in the previous two sessions. CRIC presented this scenario, generated through a synthesis of the wide-ranging discussions held within the groups. The scenario was presented by detailing the main features in each of the five sub-components, and indicating areas where there had been some inconsistency or disagreement. The 'workshop scenario' was then passed back to participants to consider, in groups, the following:

- which features do you agree with?
- where you do not agree with a particular feature, why?

Identifying Potential Areas for DTI Support

Participants were asked generate ideas for DTI support by considering six related issues:

1. Is there a role for DTI in creating a UK Flagship Centre OR should DTI support a more widely distributed capability?
2. What should be the balance between infrastructure support (computing, databases) and research support?
3. How should DTI support be integrated with research council and charity support?
4. How should the DTI stimulate university – industry interaction?
5. How could DTI support bioinformatic skill formation and mobility?
6. What should be the relationship between DTI and EU support? E.g. how closely should DTI be meshing with EU frameworks?

Each group was given the opportunity to consider each of the questions.

Prioritising DTI Support

In this session, participants were asked to vote on which of the possible DTI measures they support. Each participant was given five 'most favoured' votes, and five 'least favoured' votes to indicate their support. They were entitled to cast these votes as they wished, with the possibility for using all on one initiative or spreading the votes amongst several alternatives.

6.2 The Workshop Consensus Scenario

In this section, the workshop scenario that emerged from the consensus forming process and the synthetic analysis of the results is presented. A wide basis of agreement was arrived at which is expressed in the summary of the scenario below. This can, broadly, be understood as the vision of a realistic and credible future for UK bioinformatics in five years time. The scenario presents the five interrelated dimensions of the future: science and technology activity, institutional settings, resource flows combined here with economies of knowledge, and skills restructuring.

The scenario has various striking features, and should be seen as the outcome of the different viewpoints, individual and organisational, brought to the process. A different grouping of people would no doubt have produced a different scenario. The scenario as presented back to the workshop, with the endorsements and expressions of dissent, is reproduced in Appendix B.

Science and technology

The vision of the future had two strong emphases. Firstly, there has been integration across the spectrum of informatics (bio-, chemo-, demo-, enviro-), and within the 'bio-' from molecular to organism scales. Secondly, interoperability had been advanced by the establishment of quality standards, which had improved both data input and annotation. In terms of hardware, interoperability was considered primarily in terms of broad bandwidth internet based systems. In addition, different modelling and analytical techniques drawn from diverse disciplinary domains have produced significant scientific advances. Mathematical and statistical (e.g. Bayesian) techniques have developed within bioinformatics.

Institutions

Institutionally, UK bioinformatic activity is pivoted around a central hub – the Hinxton campus – but there were also strong views expressed and recorded that there was a need for other interconnected centres of excellence. This point is further elaborated below in relation to funding strategies. There was also a strong view that bioinformatics is essentially international, not bounded by regional or geo-political contexts, possibly reflecting a view about the 'universality' of science. So, there is no strong national or regional institutional context, and some explicitly objected to a European frame of reference.

Resource flows

There has been continued strong public funding for public science institutions, this being a powerful tradition within European countries. NGOs have also maintained a high level of funding for public science activity. Some funding for pre-competitive activity, exemplified by the SNP consortium, has provided a model in limited areas. Resource flows generated by the private sector are retained within the private sector.

Skills

A start has been made to the long-term goal of re-disciplining biology, to take account of the need for mathematical skills and new forms of experimentation. This has involved changes in curriculum at the very earliest stages, right through to a restructuring of university departments. In the meantime, interdisciplinary exchanges and groupings have been fostered to bring to bioinformatics methods and theoretical tools from other science backgrounds.

Economies of knowledge

There is a strong separation of the public and private spheres: public domain generated bioinformatic knowledge is maintained in the public domain, with open access. Private domain and intellectual property was deemed appropriate only for the added value produced by commercial activities in the private sector. A strong division is maintained between a public science research agenda oriented towards fundamental science, and a commercially driven R & D development oriented towards drug discovery, diagnostics, and healthcare.

Issues Arises from Consensus Scenario

In synthesising the outputs from the various groups into this consensus scenario, it was CRIC's view that there were two particular areas where the scenario lacked consistency or coherence. Following from the consensus view of the 'separation of public and private spheres', it was felt that neither flagship UK dedicated bioinformatics firms nor major transnational pharmaceutical companies figured strongly in the scenario. So the question was put as to whether pharmaceutical companies' bioinformatics R & D would drift to the USA, and whether this in turn would present a threat to the continued development of public science bioinformatics. The main response was to consider it desirable for pharmaceuticals to have a strong presence in the UK, and that, in the absence of credible alternative policies, maintaining a strong public science bioinformatics base could continue to provide a magnet, of whatever pulling power.

The second perceived issue of inconsistency or incoherence related to the organisation of public science. It was suggested that the vision of interoperability between informatic databases across the range and scale of informatic domains implied high levels of both compute power and accessibility. Within some informatic domains (genomics, proteomics) the need for a step change in development of compute power is already perceived. Yet, the workshop had initially given little response or endorsement of the concept of GRID technology, which in turn could possibly entail a major institutional change, inasmuch as it involves the creation of a radically new computing infrastructure. A strong view emerged that there had been as yet little involvement with the bioinformatic community – potentially a critical user group – in the GRID development process. It was considered that a bio-GRID, particularly a bio-medical-GRID would need to build in specific requirements related to access and data protection. There were also diverse views, depending on perceived future needs, as to

the importance of GRID technology, with a recognition that growth in compute power is imperative in some key bioinformatic areas.

In summary, it can be seen both from the nature of the consensus and from the unresolved issues, that the scenario forming process should be seen as a starting point for a strategic visioning of the future of UK bioinformatics. The workshop scenario was an outcome of the expressed views of a number of significant players from the UK bioinformatics community rather than that of a comprehensive and balanced representative group. For that reason, it is important to think of the scenario as the useful beginning point for further consultation, rather than as a finished and refined product.

In many ways, this vision of the future builds strongly on existing strengths and institutional arrangements, rather than suggesting any radical change or departure from them. The theme of the importance of public science and its autonomy from private sector interests or involvements was strongly represented at the workshop. Indeed, a view was expressed by one group that there was already too much blurring of private and public. So, for example, the role of pharmaceuticals as 'pulling through' bioinformatics met with limited endorsement, and was felt by some groups to be too simplistic a view, and that public health care was as important a stimulant to the development of bioinformatics.

6.3 Prioritising DTI Support

This section presents the results of participants' indications of support for alternative DTI support measures for bioinformatics. The entire list of possible actions generated during the workshop is listed in Appendix C, with the associated number of 'most favoured' and least favoured' votes for each.

Given that the result of this process is highly contingent on the particular composition of representatives at the workshop, it would seem inappropriate to apportion too much weight to the precise number of votes cast, and risk giving the impression of being overly scientific. They should be seen as initial indications of support (either positive or negative), to be used as basis for further consultation. Following this, the results presented here simply show which measures received most support.

1. Large UK Bioinformatics Facility

Overall, this received the most support, and in terms of specifying the nature of this facility in more detail, two particular models emerged. Most popular was the idea of a bioinformatics facility with a prominent medical orientation, perhaps including health informatics activities in conjunction with the bioinformatics. The other idea, named the Galen or Darwin Institute by the workshop, would be a bioinformatics facility based on the Newton Institute for Mathematics (with the primary aims of promoting interdisciplinarity and interactions between private and public research).

2. Support for database maintenance / administration / curation

There was a lot of support for this, as a necessary research infrastructure, interoperability and research development. Discussions recorded during the generation of ideas noted that it is currently very difficult to secure resources for the maintenance of databases. Funds are frequently available for establishing new databases as adjuncts to research grants, but follow on administration and support for further curation is significantly lacking.

3. Tax breaks for industry-based financial support for universities

Without specifying details, it was thought that fiscal means of encouraging private investment in public science would benefit overall bioinformatic capability.

4. Two way industry – academia sabbaticals

Discussions further elaborated the idea of two-way sabbaticals, by stressing that it would be important to make this a possibility for individuals across the industrial spectrum from SMEs to large TNCs.

5. Support for post-graduate training

This was supported by participants through two alternative mechanisms: on the one hand a co-ordinated programme across several centres; alternatively, based around a single centre, namely the EBI.

The remaining range of proposals received limited support, and some were distinctly less favoured. The only option that came at all close to attracting levels of support of the five identified above, was for DTI measures aimed at stimulating industrial contributions to the development of infrastructure.

A detailed reading of the votes cast reveals some similar preferences to the workshop scenario. In particular, there appears to be significantly different views from the public science and industry communities, strongly suggesting the existence of alternative perspectives towards the development and application of bioinformatics capability.

Conclusion

This report has attempted to develop a strategic view of the development of UK bioinformatics capability in a broad context. It presents a preliminary analysis of the current landscape, from a survey of current literature and public domain information. Additionally, interviews with some key players in UK bioinformatics provided rich and diverse insights into what is happening at the leading edge of developments. These informants then became participants in a workshop whose objective was a visioning one of developing a scenario, both desirable and credible, of the shape of UK bioinformatics in five years time. This in turn produced a view of where DTI policy initiatives, including funding, could best be targeted.

The framework of analysis for both the current landscape and the future scenario involved exploring the connections between five dimensions: developments in science and technology; institutional context, including public science institutions, NGOs, transnational corporations, SMEs (dedicated biotechnology and/or bioinformatic firms), and Research Councils; resource flows and funding; skills requirements and restructuring; and economies of knowledge regarding interactions within and between public and private spheres.

In the current landscape a number of key features emerged. UK bioinformatics has a strong presence on a global stage within a number of key areas: proteomics, crystallography, and sequence databases. Public science institutions in the UK, both national and European, provide a significant platform, notably centred around the Hinxton campus, but supported by the emergence of centres of excellence that are currently being consolidated. There is a strong UK presence of major pharmaceutical companies, and key global players in agri-food genomics are located, notably around the John Innes Centre. The UK also has stimulated the growth of SMEs, especially around these two main cluster magnets. Long term strategic national funding resources, complementing funding at the European level, are needed to sustain and grow these capability resources. The changing nature of biology and the disciplinary challenge presented by bioinformatics has primarily been addressed by increasing encouragement of interdisciplinarity in a variety of ways, prior to the longer term requirement for a redisciplining of biology towards different experimental methods and mathematical, statistical, and computer modelling.

Future horizons for UK bioinformatics present important challenges, scientific, institutional, and geo-political. Perhaps the key feature is diversification: the informatics revolution in life sciences leads and feeds into increasingly different fields, as much in terms of fundamental science as in technologies and applications. There is widespread recognition that integrated informatics (bio-, medico-, chemo-, demo-, enviro-) will present significant new demands on compute power, and become an important driver for the next generation of supercomputers. GRID technology, and specifically its adaptation to bio- and medico-informatic objectives, will entail both resource and institutional change. At the same time, much significant bioinformatic development will be independent of GRID infrastructures, some of it through networked broad bandwidth internet systems. Serious consideration will be needed to

address the signs of a shift of major pharmaceutical R&D in bioinformatics to the USA, with likely consequences for the public science institutions. A key role will continue to be played by the Hinxton cluster, at national, European regional, and international levels, complemented by other centres of global excellence.

To assist in realising their scenario, the priority areas of DTI support for bioinformatics considered to be 'most favoured' by this workshop were as follows:

- Large UK Bioinformatics Facility
- Support for database maintenance / administration / curation
- Tax breaks for industry financial support for universities
- Two way industry – academia sabbaticals
- Support for post-graduate training

As a document which is the outcome of some preliminary research and an initial process of consultation, this report is designed to serve as an instrument for further consultation, from a wider constituency. The aim is to stimulate further strategic thinking about the future of bioinformatics, and to assist in the process of policy formation and funding.

Appendix A: Questions for Scenario Generation

Science and Technology

Question 1: What will be the key developments in Informatic Domains

- Within biology (genome, proteome, crystallography, enzymology, microscopy, etc.)
- Links between wet and dry science
- Integration between bio- and chemo-, demo-, enviro- etc. informatics

Question 2: What will be the key developments in informatic tools

- Software
- Pattern recognition, modelling
- Mathematics

Question 3: What will be the key developments in Infrastructure technologies

- Internet
- Interoperability

Question 4: What will be the key developments in orientation towards pharmaceutical and agri-food applications – the extent to which bioinformatics is contributing to / driving them

- Lead drug discovery, gene therapy, diagnostics
- Stress tolerance, nutrient dense foods

Skills

Question 1: What will be the key developments in the disciplining of biology at different educational stages?

- experimentation
- analysis

Question 2: What will be the key developments in interdisciplinarity?

- biology, computation, mathematics, artificial intelligence, engineering, epidemiology etc.
- modes of integration

Question 3: What will be the key developments in skill retention and recruitment?

- public sphere
- private sphere
- international migration
- sectoral migration

Institutional Organisation

Question 1: What will be the key developments in the organisation of publicly funded bioinformatic research?

- a flagship publicly funded bioinformatics hub in the UK
- the organisation and role of university bioinformatic research

Question 2: What will be the key activities of transnational corporations in UK bioinformatics?

- pharmaceutical and agrofood
- UK activities vs global activities
- linkages with UK SMEs

Question 3: What will be the key activities of UK DBIFs and Bioinformatic tools providers?

- cluster formation
- linkages with other firms

Question 4: What are the best modes of interaction between commercial and public institutions?

- US vs. European vs. UK modes

Resources and Economies of Knowledge

Question 1: What are the key sources of Public and NGO investment?

- European
- UK
- governmental and research council

Question 2: What are the key sources of private investment?

- venture capital and stock market
- TNC investment
- mergers and acquisition

Question 3: How is intellectual property protected?

- private (e.g. patents, secrecy, firewalls etc.)
- public (open and confidential databases)

Question 4: What is the nature of flows of knowledge between public and private institutions?

- data, information, algorithms etc.
- people

Appendix B: The Workshop Consensus Scenario

Science and Technology	
✓✓D C	1. integrated informatics: quality (standard); from molecule to population (domain); sequence, text, image..... (type)
✓✓✓ ✓ _{DC}	2. Broad bandwidth internet and interoperability solutions
✓✓✓D	3. Different modelling and analytical techniques interacting with different disciplines in different domains
✓ _x C	4. Bioinformatics pulled through by pharmacogenomics and healthcare
Institutions	
✓✓✓ ✓ _{DC}	1. At least one hub (with major institutionalised databases) with strong complementary interdisciplinary research platforms
	2. Good business models for SMEs to act as magnet for pharma TNCs, encouraged by strong European connections
✓✓✓D C	3. Encourage more joint public-private pre-competitive research consortia
Resources	
✓✓✓ ✓ _D	1. Large scale and sustained infrastructure funding, particularly for databases, with problem oriented facilities
✓✓ _D	2. Hybrid funding for pre-competitive research
✓✓✓ ✓ _{DC}	3. Public domain (<i>funded</i>) generated data is open - IP on Know-how to produce value added products
Skills	
✓✓ _x ✓ D	1. Re-disciplining biology throughout all levels of education (with institutional restructuring across university departments, and building research careers)
✓ _x D C	2. Establishing a national training and technology transfer centre promoting two-way interdisciplinarity
✓✓✓D	3. Strategy for skills provision addressing issues of recruitment and retention.

The Workshop Scenario.

In the left hand column symbols represent views about the consensus scenario expressed by subgroups: ticks represent endorsement; C represents a credible but not necessarily desirable outcome; D represents a desirable but not necessarily credible outcome; x represents disagreement.

Appendix C: Voting for DTI Support

What is most/least important for the DTI to support either alone or together with other funding bodies?

Each participant was given 5 red and 5 green dots

Green dots – most important

Red dots – least important

Balance between Infrastructure / Research

	Green Dots	Red Dots
Support for database maintenance / administration / curation	21	0
DTI facilitating / seeding industrial support for infrastructure	6	1
DTI to fund start-ups of SMEs	1	11
DTI support for buildings, equipment, technical support staff	1	8

Relationship EU-DTI support

Mentoring / practical support in developing practical proposals i.e. that “play the game”	0	1
Support for SME focus to industry programme at EBI / other EU hubs	4	0

DTI support to be integrated with Research Councils?

Cross-disciplinary “town meetings” (With DTI / Research Council presence)	3	5
Information exchange mechanisms	0	2
Research / business interface	0	0
Multiple SMART awards held simultaneously	2	2
DTI seed funding for new companies	4	5

Skills and Mobility

Promotion of centres of excellence for skill formation – business orientated	1	0
Two-way sabbaticals	11	2
Educate venture capitalists	0	6
MBA Fellowships in biology	1	11

Large Facility

Large Facility	4	0
Cauldron	0	1
Darwin / Galen – bio equivalent of Newton	15	7
Health informatics Remember the need to add genetic data to health records – so it's not just health informatics <i>Consider making use of existing health informatics provision rather than duplication.</i>	17	0
Distributed <i>Coordinated distributed centres around the country Coordinated distributed network</i>	0	2
Rotated	0	12
TOTAL FOR LARGE FACILITY	36	21

UK Flagship or distributed?

Strong national centres in each sub-domain	4	5
--	---	---

University / Industry interaction

Transform and extend HGMP-RC into national re disciplining / training centre for industry and academics	2	9
Problem-oriented rotating workshop	2	1
Professional informatic chartered association	0	13
Training for all levels of post-grad at Hinxton (<i>aka EBI</i>)	4	3
Support existing industry – university centres of excellence	1	0
Matching funding for masters courses	0	2
Post grad training in several centres, in coordinated programme	7	0
Tax breaks for industry financial support for universities	14	0
Marriage-broking for industry to academics	0	10

Appendix D. Bioinformatics Workshop Participants July 2001, Warwick Campus

Participants

Industry

Dr Mark Swindells - Inpharmatica
Dr Tom Flores* - Synomics
Dr David Parry-Smith - Biofocus plc
Dr Charlie Hodgman - GSK
Dr Jerry Lanfear* - Pfizer

Dr Susie Stephens - Sun Microsystems
Dr David Hodgkinson - Quintessa
Mr David Wallder – Korn/Ferry Int
Dr Simon Brocklehurst - CAT
Dr Michael Cole - Amersham
Prof Nick La Thangue* - Prolifix
Mr Will Dracup - NonLinear Dynamics
Dr Chris Rawlings - Oxagen

Academia

Prof Terri Attwood - Manchester
Prof Dave S Broomhead - UMIST
Prof Steve Muggleton - Imperial
Dr Olaf Wolkenhauer - UMIST
Dr Chris Ponting - MRC Functional Genetics Unit, Oxford
Prof Michael Ashburner - EBI
Dr Alan Robinson* - EBI
Dr Richard Durbin - Sanger
Dr Richard Baldock* - MRC HGU
Prof Carole Goble - Manchester
Dr Peter Artymiuk - Sheffield
Prof David Willshaw - Edinburgh
Prof David Ingram - UCL

Funding Bodies

Dr Colin Miles - BBSRC
Dr Lesley Thompson - EPSRC
Dr Diane McLaren - MRC
Dr Alan Doyle - Wellcome

Dr Celia Caulcott - Applied Genomics LINK programme co-ordinator

Dr Monica Darnbrough - Director, Biotechnology Directorate, DTI
Alisdair Wotherspoon - Head, Knowledge Transfer Team, Biotechnology Directorate, DTI
Dr Norman Freshney – Office of Science and Technology, Science and Engineering Base

Dr Andrew McMeekin - ESRC Centre for Research in Innovation and Competition (CRIC)
Dr Mark Harvey - ESRC CRIC

Ms Helena Poldervaart and Ms Lynn Wetenhall - Projects in Partnership (workshop facilitators)

(Paula Milton - LGC - venue liaison)

* denotes unable to attend workshop, although contributed to the preparation of the workshop report

References

- Attwood, T. K. and Parry-Smith, D.J. 1999. *Introduction to Bioinformatics*. Prentice Hall. Harlow.
- Attwood, T.K. and Miller, C.J. 2001 'Which craft is best in bioinformatics?' *Computers and Chemistry*.
- Botstein, D. 1999. *The Biomedical Information Science and Technology Initiative*. National Institute of Health Report, June.
<http://www.nih.gov/about/director/060399.htm>
- Biotechnology Strategic Forum, 1997. 'Financing Biotechnology Databases' Workshop, Purmerend, The Netherlands. <http://btsf.embnet.org>
- Biotechnology Strategic Forum, 1998 'Building and Owning Biotechnology Databases' Workshop, Purmerend, The Netherlands. <http://btsf.embnet.org>
- Brown, N, Nelis, A, Rappert, B., Webster, A., van Ommen G.J.B. 1999. *Bioinformatics. A technology assessment of recent developments in bioinformatics and related areas of research and development including high throughput screening and combinatorial chemistry*. Final Report for the Science and Technology Options Assessment, European Parliament. May.
- Butler, D. 2001. 'Are you ready for the revolution?', *Nature*, 409, 758-60.
- Coleman, L. 1998. 'Directive 96/9/EC on the legal protection of databases.' *Biotechnology Information Strategic Forum*. The Netherlands.
<http://btsf.embnet.org>.
- Ellis, L.B.M. and Kalumbi, d. 1999. 'Financing a future for public biological data.' *Bioinformatics*, 15, 9, 717-722.
- Ellis, L.B.M. and Attwood, T. 2001. 'Molecular Biology Databases: today and tomorrow.' *Drug Discovery Today*, 6, 10, 509-13.
- EMBL (European Molecular Biology Laboratory). 2001. 'EU Funding Boosts Bioinformatics Infrastructure in the Post-genomic Era'. Press release.
<http://www.ebi.ac.uk>.
- Gavaghan, H. 2000. 'Training: Europe seeks solution to bioinformatics shortfall', *Nature*, 404, 687-8.
- Grindrod, P. 2001. *UK Bioinformatics for Functional Genomics: Watching the Detectives*. Numbercraft.
- Hasty, J., McMillen, D., Isaacs, F and Collins, J.J. 2001. 'Computational Studies of Gene Regulatory Networks: *In Numero* Molecular Biology.' *Nature*, April, 2, 268-279
- Joint Research Council Report. 2000. *Promoting Interdisciplinary research and training*. See also Joint Research Council Response.
- Medical Research Council 2001. 'Human tissue and biological samples for research: operational and ethical guidelines' MRC Ethics Series
- Moore, S.K. 2001. 'Making Chips to Probe Genes' *IEEE Spectrum*, March.
- Muggleton, S. 1999. 'Scientific Knowledge Discovery using Inductive Logic Programming.' *Communications of the ACM*, 42(11):42-46, November.
- National Research Council, 1997. *Bits of Power. Issues in Global Access to Scientific Data*. Report. <http://stills.nap.edu/html/BitsOfPower/>
- Powledge, T. M. 2001. 'Changing the rules? The agreement between Celera and Science.' *EMBO Reports*, 2, 3, 171-2.

- Reichardt, T. 1999. 'It's sink or swim as the tidal wave of data approaches.' *Nature*, 399, 517-520.
- Rybak, B. 1968. *Psyché, Soma, Germen*. Gallimard. Paris.
- Rybak, B. 1978. 'Bio-Informatics and Bio-Process in the Physiology of Communication.' *Biosciences Communications*, 4, 3-4, 158-9.
- Uhlir, P.F. 1998. 'Potential impacts on research from proposed US Database IPR legislation.' *Biotechnology Information Strategic Forum*. The Netherlands. <http://btsf.embnet.org>.
- UKHEC, 2000. *A Review of UK HEC Grid Infrastructure. State of the Art and Next Steps*. <http://www.dl.ac.uk/TCSC/UKHEC/GridWG/EPSRCGrid.pdf>
- Wolkenhauer, O., 2001a, *Data Engineering*, Wiley
- Wolkenhauer, O., 2001b, 'Systems Biology: The reincarnation of systems theory applied in biology?', *Briefings in Bioinformatics*, vol. 2, no. 3.