e-Science and e-Infrastructure needs of UK Life Sciences industries.

A report for the UK e-infrastructure Leadership Council

Darren Green, Douglas Kell, Ian Dix, Anne-Marie Coriat and Louise Leong

Executive Summary

All of the developed economies of the world are looking to grow their Life Sciences sector. The sector typifies the new knowledge economy that governments are seeking to promote, with enormous potential for the growth of new businesses and high-waged jobs. Recently the Prime Minister (5th December 2011) summarised the importance of the sector to the UK:

"The UK life science industry is one of the world leaders; it is the third largest contributor to economic growth in the UK with more than 4,000 companies, employing around 160,000 people and with a total annual turnover of over £50 billion. Its success is key to future economic growth and to our goal to rebalance the economy towards making new products and selling them to the world. Globally the industry is changing with more focus on collaboration, out-sourcing of research and earlier clinical trials with patients."

The Life Sciences sector is on the brink of a data deluge, as for example, genome sequencing becomes a commodity and literature publication rates spiral (2 peer-review papers published per minute in bio-medicine alone). The draft sequence of the human genome took scores of labs some 10y and cost ca $3 Billion. With current experimental equipment this can be accomplished in less than a day, and the cost of obtaining an individual genome sequence is set to become just a few hundred pounds. The combination of such massive amounts of literature data and knowledge, typically available in digital form and online, means that a competitive Life Sciences sector must be well placed to harvest and exploit these wonderful resources.

For the Pharmaceutical industry, the increasing availability of electronic health records is set to enable the use of more effective treatments, improvements in drug safety, assessment of risks to public health and identification of the causes of diseases and disability at a speed and scale not previously possible. It is vital that the UK research community is in a strong position to maximise the research potential offered by linking electronic NHS records with other forms of routinely collected data and research datasets, such as that from the UK Biobank. The MRC led a mapping exercise on behalf of the UK research funders and the Association of British Pharmaceutical Industries (ABPI) in 2010 to review the existing UK capability and examine the requirements to support a sustainable research base in the future. The report highlighted the need to build capability and capacity in health informatics research and for further methodological development in the use of electronic records for research, including complex data linkage.

Other nations are actively investing in e-science. For example, the USA has recently announced an additional $200 million of funding for a "Big Data" research and development initiative, whilst the German government has been conducting a similar consultation. A common theme is the need for software, sustainability of data resources and the development of a skilled workforce. These conclusions are echoed in this report, that builds on a variety of others such as those at http://www.rcuk.ac.uk/research/xrcprogrammes/OtherProgs/eInfrastructure/Pages/home.aspx.

The UK has particular strengths that should be built upon. The research base has responded to the pending data deluge, with BBSRC, MRC and NERC all setting up centres. The Computational Genome Analysis and training programme (CGAT; see http://cgat.org/cgat/) is a 5-year strategic training

award by the MRC Strategy Board. The programme helps to address the UK-wide shortage of computational biologists capable of analysing and interpreting next-generation sequencing data sets. The programme trains post-doctoral researchers in next-generation sequencing analyses while at the same time providing analytical capacity to UK-based experimental groups. BBSRC's response to the sequencing revolution and its attendant data deluge was to set up The Genome Analysis Centre (TGAC) on the Norwich Research Park. In addition to the investments by research councils, the UK is fortunate to be the home to the European Bioinformatics Institute, a world leader in the application of e-science. Recently, the UK announced a £75 million investment to help establish the ELIXIR hub (European Life Sciences Infrastructure For Biological Information).

The pharmaceutical, agrochemical and animal science industries are a traditional strength of the UK. In spite of recent downsizing, there remains a high quality pharmaceutical sector, both large companies and biotech, with strong computational functions. Although much of the investments described above are aimed at Biology, Chemistry is of equal importance to industry, being the primary generator of intellectual property. There are several outstanding computational chemistry and chemoinformatics research groups in UK universities with good links to industry and The Royal Society of Chemistry is an international leader in chemical publications. It is important to industry that this research base is maintained and enhanced.

The UK has made significant investments in High Performance Computing which has resulted in pockets of significant capability and expertise. Whilst the quality of the resulting infrastructure is not in question, it is important to acknowledge that previous investments in the UK e-infrastructure have not resulted in large scale adoption by the UK Life Sciences industry. Therefore, it is essential to consider how any new government investment might be exploited by industry, by focussing investment where it will be used most, by new collaboration models and by easier access to the academic institutions that receive the investments. The focus of the e-infrastructure investment has been commented on by a recent Royal Society report, *Science as an open enterprise* :

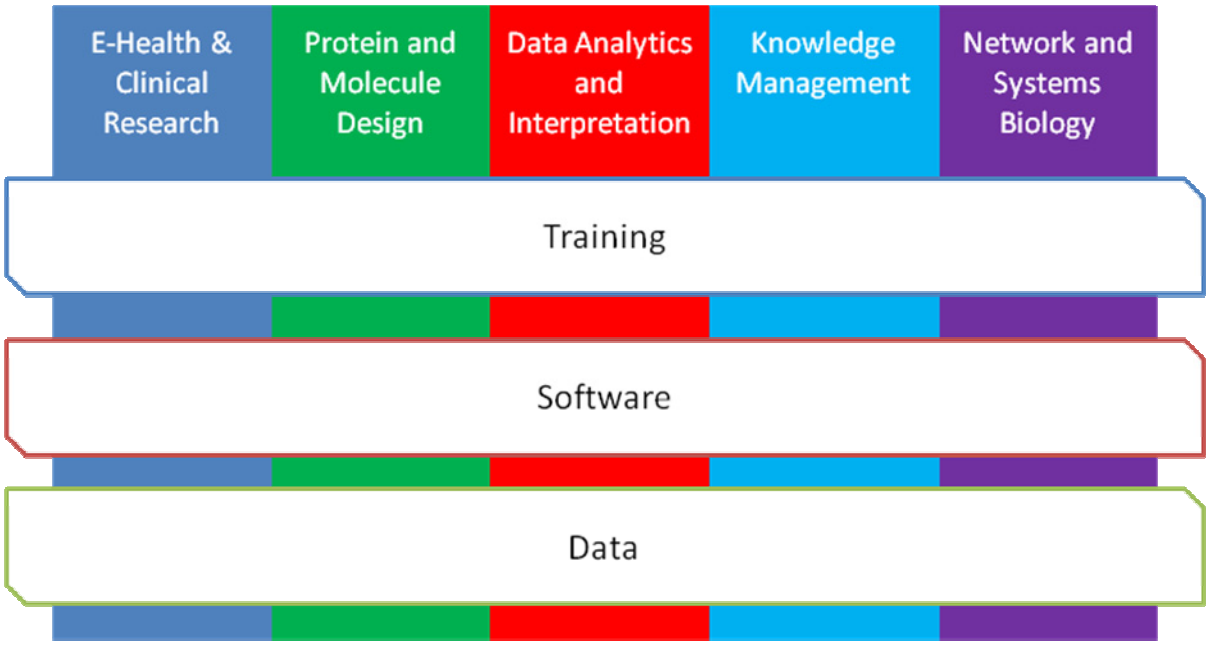
"The UK government, through the Department of Business, Innovation and Skills, should revisit the work behind its roadmap for e-infrastructure. The urgent need for software tools and data scientists identified in this report need to be given due prominence alongside the physical infrastructure, skills and tools needed to exploit the data revolution. It should consider a major investment in these areas ensuring that the UK is able to exploit the data deluge".

Given the desire for greater application of e-science in industry and the significant investments already made by government, what is preventing the required uptake? We recognise several barriers to progress:

1. For e-health applications, there is not yet a national infrastructure to ensure ethical, secure access to appropriate patient data and to enable collaboration between academics, industry and NHS clinicians.
2. There is no overarching strategic vision from industry, HEFCs or the research councils which can be used to plan investment.
3. Central investment in high performance computing has focussed on limited hardware and software applications. Life Sciences have a set of diverse needs and this has resulted in a fragmentation of the UK infrastructure.

4. Previous investment has been primarily in physical capital. The Life Science sector places greater emphasis on Human Capital – scientists who can manage and interpret the coming deluge of data – and sustainable software development for the many informatics tasks.
5. The UK Pharmaceutical and Agrochemicals sectors have been significantly downsized in recent years. The commercial environment for both sectors is challenging and this limits the potential for industrial investment in long term infrastructure.
6. Computational science has been added to lab-based science departments in Universities, which have traditionally not integrated well. As such, there are no "e-science" departments offering undergraduate (or even postgraduate) degrees.
7. Lack of career paths for scientific software developers and database curators in academia.
8. Pressure in industry mean that the traditional routes for academic collaboration – CASE awards or post-doctoral funding – are either too long term or too expensive. More immediate access to academic expertise for shorter periods is required.

What is termed the Life Sciences sector is actually a highly diverse collection of disciplines and applications. It is helpful to categorise our recommendations not by sector but by activity, and to consider the e-infrastructure requirements for each activity as summarised in the graphic below:



There is a consistent recommendation which is common across all the activities: **we must invest in the UK skills base through better training.**

There is a need for increased skills in health informatics, systems biology, Protein and Molecule design, librarianship, legal compliance, information management, text analytics, semantics & semantic technology, social media analytics and visualisation.

We recommend that

- Multiple specialist Masters level conversion courses are established, appropriate to the sectors
- Shorter courses are made available for Continuing Professional Development of industrial scientists.

- A career structure is developed in academia for enabling roles such as data managers, software engineers, informaticians and data analysts
- E- Science modules should be incorporated into all Life Science undergraduate degree courses

For each activity there are also specific recommendations.

Health and Clinical Research

Software

- A priority is the development of a system to provide a single access portal for research studies in the UK.
- The UK should develop an infrastructure that enables academics, hospitals and industry to collaborate in an environment that realizes the required ethical standards, privacy policies and data access rights. At the very least, standardised dictionaries, ontologies and software interfaces should be agreed on and consistently applied.
- Further investment is needed in methodological research for complex data linkage.

Data

- NHS Trusts need to prioritise the provision of dedicated funds for E-health records systems with high quality data (this should be a responsibility of the Trust). The National Institute for Health Research, and devolved administrations, should further develop investments in Biomedical Research Centres and Biomedical Research Units to generate a rich UK wide data set.
- Deliver current investments in data services such as CPRD and SHIP.
- Targeted training/engagement of healthcare professionals is needed to encourage use of electronic health records.

Network and systems biology

Software

- A priority is investment in the development of the tools of network and systems biology that are usable by working biologists without a computer science background.

Data.

- Support the Open Data agenda together with suitable interoperable data models that will allow the federated delivery of data describing biochemical systems and that then allow integration with our literature knowledge to provide semantically annotated models of biological systems.

Protein and Molecule Design

Training

- Investment in hardware and computer networks needs to be suitably complemented by investment in people with life sciences domain knowledge. All e-infrastructure should have people dedicated to facilitate training of scientists and enable access to the appropriate high performance computing resources for their scientific problem (the “On Ramp”).
- Consider a Knowledge Transfer Network dedicated to e-Science.

Software

- The e-infrastructure should host a mixture of commercial and open source software. This will reduce duplication of effort in academia and at the same time increase robustness and supportability for industry. Create at least one, preferably more, Centres of Excellence for Molecular Design in the UK, which will produce a critical mass of researchers to pioneer new

Data

- The e-infrastructure must be designed and implemented such that proprietary data can be routed and stored securely. Without complete confidence in the fidelity of the infrastructure it will not be possible to attract industrial use.

Data Analytics and Interpretation

Software

- In order to be competitive the UK must invest in the development of novel algorithms. This will drive improvements in data analytics, machine learning, advanced statistics and visualisation.

Data

- While data science assumes (and requires) the availability of the relevant data, it will be important that the relevant data are in fact available (the Open Data agenda) together with the necessary metadata set out according to recognised standards of interoperability. Assistance for developing, and where necessary imposing, such standards will be required.

Knowledge Management

Software & Data

- The UK should provide core Knowledge Management services over UK generated content and beyond. Immediate opportunities should centre around the concept of 'The UK Book of Science' covering all UK funded reports, scientific papers, grants, patents, conferences etc. Initial services could include 1) UK Key Opinion Leader analysis enabling users to find out who is working on what innovation enabling UK based collaborations or 2) Derived Knowledge Services allowing users to rapidly navigate automatically generated knowledge summaries for any scientific concept(s) to support hypothesis generation.
  Such KM services would require a maintained core compute infrastructure, federation and integration technologies, standards agreement and implementation and copyright & content access agreement.

Contents

1. Introduction

All of the developed economies of the world are looking to grow their Life Sciences sector. The sector typifies the new knowledge economy that governments are seeking to promote with enormous potential for the growth of new businesses and high-waged jobs. According to a speech by Prime Minister David Cameron on the 5th December 2011 (when he announced a multi-million pound package of support for the UK Life Sciences industry).

"The UK life science industry is one of the world leaders; it is the third largest contributor to economic growth in the UK with more than 4,000 companies, employing around 160,000 people and with a total annual turnover of over £50 billion. Its success is key to future economic growth and to our goal to rebalance the economy towards making new products and selling them to the world. Globally the industry is changing with more focus on collaboration, out-sourcing of research and earlier clinical trials with patients." (Source http://mediacentre.dh.gov.uk/2011/12/05/govt-boost-to-uk-life-science-industry/)

The life sciences sector is an ideal area for the government to stimulate growth through investment and as recently as 24th May Minister of Universities and Science David Willets announced fresh funding of £250M to stimulate the life sciences industry in Scotland. (Scotland's life sciences industry employs more than 32,500 and is worth £3.1 billion a year to the nation's economy, according to Scottish Enterprise). Source http://www.scotsman.com/business/technology/scots-to-benefit-from-life-sciences-250m-funding-1-2313136

As Leroy Hood, founder of the Seattle-based Institute for Systems Biology, is wont to say "Biomedicine is an informational science". The combination of massive amounts of literature (2 peer-review papers published per minute at PubMed/Medline alone (Hull *et al*., 2008)), data and knowledge, typically available in digital form and online, means that a competitive Life Sciences sector must be well placed to harvest and exploit these wonderful resources. To take just a single example, we consider genomics data.

As is well known, the speed and cheapness of genome sequencing have increased supra-exponentially. The draft sequence of the 3Gbase human genome took scores of labs some 10y and cost ca $3Bn ($1000/kbase). Present numbers for an Illumina Hiseq exceed 200 Gbase/week and will by the end of the year be 1 Tbase/week at a cost of $0.02/kbase. The recent Oxford Nanopore Technology (ONT) announcements (http://www.nanoporetech.com/news/press-releases/view/39) dwarf even these numbers. Figure 1. Is a typical graph from a 2009 survey.

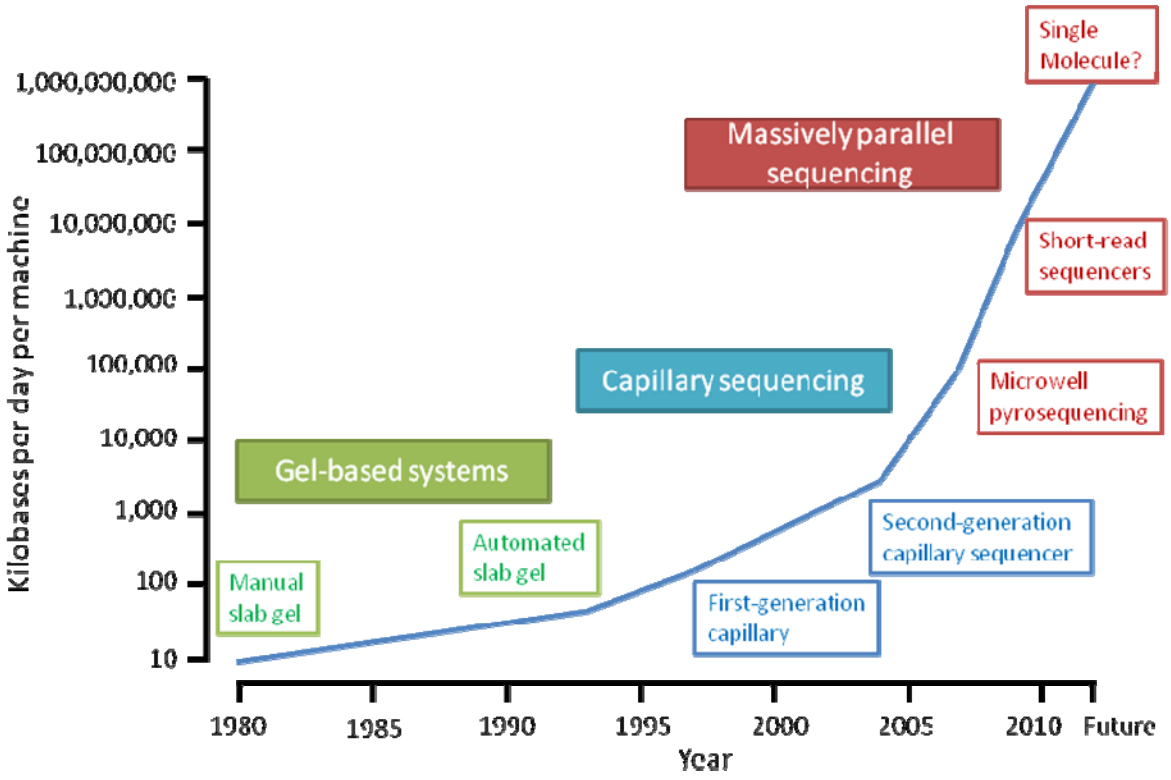


Figure 1. Projected growth in data produced from genome sequencing (adapted from Stratton *et al*., 2009).

The research base has responded, with BBSRC, MRC and NERC all setting up centres. The Computational Genome Analysis and training programme (CGAT; see http://cgat.org/cgat/) is a 5-year strategic training award by the MRC Strategy Board. The programme helps to address the UK-wide shortage of computational biologists capable of analysing and interpreting next-generation sequencing data sets. The programme trains post-doctoral researchers in next-generation sequencing analyses while at the same time providing analytical capacity to UK-based experimental groups. BBSRC's response to the sequencing revolution was to set up The Genome Analysis Centre (TGAC; http://www.tgac.ac.uk/) on the Norwich Research Park. As high-throughput benchtop sequencing becomes available to every laboratory, it becomes clear that the challenges are mainly in information handling and the development of understanding through application of algorithms, scientific knowledge or intuition. Such a centre is best seen as an '**e-science or e-biotechnology centre with some sequencing capability**'.

Note too that 'sequencing' is not just genome sequencing; sequencing of RNA, and of the products of competition experiments using libraries of 'interfering' RNA molecules, can provide swift and powerful routes to the analysis of gene function.

In addition to the investments by research councils, the UK is fortunate to be the home to the European Bioinformatics Institute, a world leader in the application of e-science. Recently, the UK announced an £85 million investment to help establish the ELIXIR hub (European Life Sciences Infrastructure For Biological Information; http://www.elixir-europe.org/about):

"ELIXIR unites Europe's leading life science organisations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. It is a pan-European research infrastructure for biological information. ELIXIR will provide the facilities necessary for life science researchers - from bench biologists to cheminformaticians - to make the most of our rapidly growing store of information about living systems, which is the foundation on which our understanding of life is built." Professor Janet Thornton, Director of the EMBL-European Bioinformatics Institute.

In contrast to the significant public investments described above, the UK Life Sciences industry has experienced significant downsizing in recent years, in particular the Pharmaceuticals sector has seen the closure of 5 major R&D centres. One consequence of this for e-science departments in industry is that the majority of staff are directed to mission-critical projects, and less are available to maintain skills in fast developing areas such as High Performance Computing. When coupled with reduced access to collaborations funding for infrastructure research (another result of the downsizing), it can be difficult to maintain contact with academic expertise or exploit the investments being made by government. And to compound this, companies are less able to hire new employees to bring knowledge of new technologies.

It is important to acknowledge that previous investments in the UK e-infrastructure have not resulted in large scale adoption by the UK Life Sciences industry:

"In the domain of high performance computing for life sciences, the Science and Technology Facilities council (STFC) runs an e-science project with a 10-year history. We are not aware of any life science company that makes use of these resources…" (Harland, 2012).

Therefore, it is essential to consider how any new government investment might be exploited by industry, by focussing investment where it will be used most, by new collaboration models and easier access to the academic institutions which receive the investments. The focus of the e-infrastructure investment has been commented on by a recent Royal Society report, *Science as an open enterprise* (see http://royalsociety.org/policy/projects/science-public-enterprise/report/):

"The UK government, through the Department of Business, Innovation and Skills, should revisit the work behind its roadmap for e-infrastructure. The urgent need for software tools and data scientists identified in this report need to be given due prominence alongside the physical infrastructure, skills and tools needed to exploit the data revolution. It should consider a major investment in these areas ensuring that the UK is able to exploit the data deluge".

Other nations are actively investing in e-science. For example, the USA has recently announced an additional $200 million of funding for a "Big Data" research and development initiative (http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf) to

- *Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data.*
- *Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning; and*
- *Expand the workforce needed to develop and use Big Data technologies.*

The Lifesciences component is led by the National Science Foundation and the National Institutes of Health and will advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large and diverse data sets. NIH is particularly interested in imaging, molecular, cellular, electrophysiological, chemical, behavioral, epidemiological, clinical, and other data sets related to health and disease.

In addition to its funding of the Big Data solicitation, NSF is also investing to develop interdisciplinary graduate programs, fund major research projects in machine learning, crowd sourcing and geosciences, create a dedicated research training group to educate undergraduates in visualisation techniques for complex data and convene researchers across disciplines to determine how big data can transform teaching and learning.

The German government has been conducting a similar consultation, and its recommendations for bioinformatics are now available (http://biooekonomierat.acatech.de/files/downloads/boer_broschueren_eng/boer_broschuere_bioinformatik_eng.pdf). Recommendations include:

- Creation of a "Comprehensive Coordination Body" to act as the coordination, contact, and information interface between a network of localised expertise centres, biology and bioinformatics research institutions, and other users and interest groups. This body would also oversee the development of standard operating procedures, uniform interfaces, and conscientious data documentation by the excellence centres
- Development of long term strategies for research, action and funding, to increase the number of collaborative public-private projects and ensure the sustainability of data resources.

The themes of coordination, data standards and sustainability are echoed in this report.

The Life Sciences industry has a diversity of applications in e-science, and requires a heterogeneous infrastructure to support it as will be illustrated in this report. To aid navigation of this diversity, the report has been organized by discipline.

2. Translational Medicine

Translational medicine has been used in recent years to characterise biomedical research that aims to translate between Clinical Practice and Laboratory research. In simple cases this involved performing molecular profiling of tissue samples collected from patients, but can extend to an extensive exploration of the biological similarities between Human subjects and their *in vivo* and *in vitro* models of disease. The complexity of the study and data integration challenge is dependent on the research/clinical question being addressed. Most translational studies are focused on the identification and validation of biomarkers that are testable in patients, including markers that are predictive of:

- the prognosis of disease (severity)
- how well a patient may respond to a pharmacological therapy
- the susceptibility of a patient to side effects of therapeutic intervention
- the identification of subgroups that are at increased risk for disease

However biomarkers are extremely broad in nature being defined as 'a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention'. Typically such characteristics can include genetic changes, antibody levels, biopsy histopathology observations, blood protein levels, blood chemistry, macro physiological observations (such as grip strength or lung peak flow) or more complex markers based on multiple observations such as Hy's Law for liver safety in clinical trials (see http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM174090.pdf). This heterogeneous nature of biomarkers creates a significant data capture, management and analysis challenge for translational medicine systems. Of particular challenge is the high dimensionality and scale of the molecular marker data, requiring advanced analytics to identify, and statistically validate, biomarkers in often very variable populations. Such molecular profiling techniques can include the high dimensional quantitative profiling of panels of mRNA, Proteins, Lipids, Metabolites and Antibodies and the identification of genetic variations between patients (copy number, SNPs, fusions, methylations etc) through the unprecedented advances in sequencing technologies ("Next Generation Sequencing (NGS)"), enabling sub-$1000 genome sequencing.

Notwithstanding the complexities of the area, there are exciting possibilities for patient care should we be able to make progress. For example, a clinical trials analyst may ask the question "*How many patients with Breast Cancer have a smoking habit and an Erythrocyte count below 5? Of those patients, which of them have provided samples and what is the differential expression of genes in those samples compared to a publicly available control set. What common cellular processes are associated with these differentially expressed genes? ".* To answer this question requires integration of clinical, tissue sample, genomic data (public and private) as well as reference knowledge of gene and molecular signalling pathways. As well as being a significant data integration and analytics challenge, access to patient data raises clinical, ethical and political issues, solutions to which must be reflected in any translational e-infrastructure.

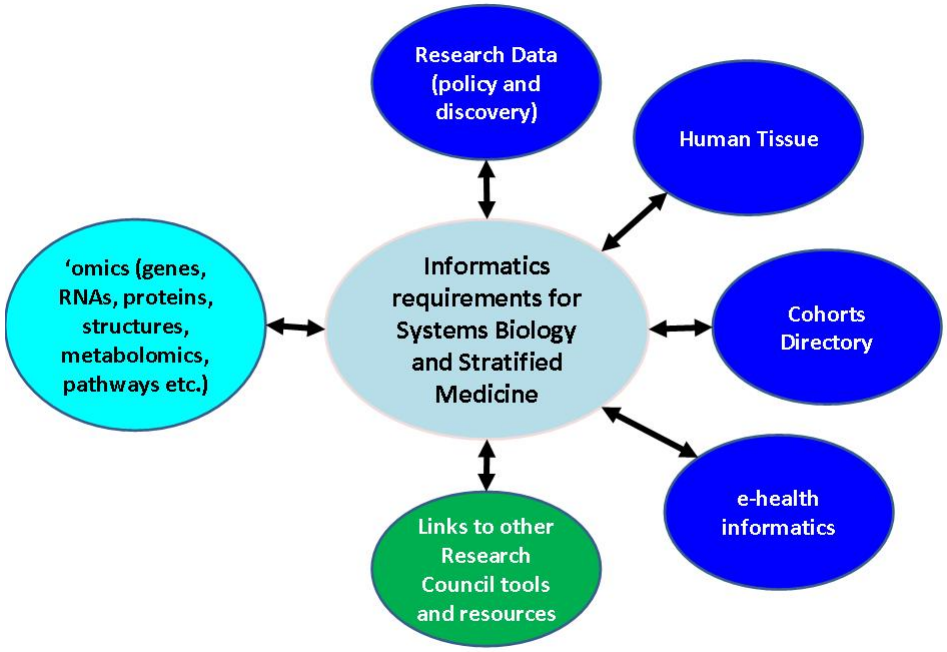
As a result many Pharmaceutical companies have complex internal Translational Research (TR) and Knowledge Management (KM) processes, standards and services in place. However the Pharmaceutical business model is rapidly changing from fully integrated, siloed, R&D organisations to fully networked organisations with R&D conducted across multiple independent commercial, academic, governmental etc organisations. This change has been driven in part by the rapid advances and understanding of human pathophysiology and pharmacology requiring different specialists, such as Clinical Scientists, Laboratory Scientists, Information Scientists, Statisticians and Heath Economists (etc), to come together in dynamic, temporary collaborations. Each of the specialisms listed above has their own underlying set of technology infrastructure to facilitate the collection, organisation, integration and analysis of data, all of which need integrating to support translational analytics. Examples of some of these capabilities are described in Table 1.

Table 1 – A summary of the Baseline IT platforms to support Translational Research

Scientific Discipline	Infrastructure Components
Clinical Sciences	• Document Management to manage trial approval and patient consent forms • Electronic Case Report Form (eCRF) data collection system • Clinical Data Management platforms • Clinical Statistics Platforms • Medical History records (eHRs)
Biobank	• Document Management to manage trial approval and patient consent forms • Laboratory Information Management Systems (LIMS) for tracking the location of samples
Biological Sciences (Bench)	• Electronic Notebooks to capture specific experiments •
Biological Sciences (High Dimensional Biology)	• LIMS systems to organise workflow and capture results files • Data Storage Archives to store large primary data files from analytical platforms (imaging, NGS, omics, etc)
Biostatistics/Bioinformatics	• Statistical/Data programming environments for processing and analysing data • Reference Databases of biological information
Knowledge Management/Systems Biology	• KM tools to capture results and output of all experiments • Modelling tools to combine data from all domains for analysis • Reference knowledge (literature, pathway knowledge, etc)

This shift requires a new generation of TR KM infrastructures that enable cross-organisational, collaborative, secure working between Pharma, Biotech, Academic, Medical, Patient Association and even Regulatory partners, in temporary project teams around common translational problems. The challenge we have is that currently such infrastructures either do not exist or are being built to meet the needs of individual consortia. What is required is a re-useable, secure TR KM infrastructure service and components that can be rapidly re-deployed and configured for such cross-organisational investigations. The key features of such a platform include: Multi-terabytes of storage, rigorous access control (critical in handling patient data), data governance and curation services, standardised dictionaries, ontologies and APIs (Application Programming Interface), ETL (Extract Transform Load) tools to carry out loading of data, high bandwidth connections to data provision centres, data modules enabling the management of a wide range of data modalities,

patient and sample level data tracking (enabling data retraction), overlaid collaborative search and analytics tools, virtual team collaboration spaces, all of which are available as a sustainable service which can either host multiple collaborations or be flexibly deployed to meet the needs of specific collaborations. On top of this such an infrastructure needs secure connections with medical eHR systems, biobanks and Laboratory Information Management (LIMS) systems. These requirements and interdependencies have been summarised in a study by MRC (see www.mrc.ac.uk/e-health):



Strategic discussions are ongoing within the NHS about approaches to maximise data linkage between laboratory and clinical or population data, in part through the government's Human Genomics Strategy Group (HGSG), which is exploring skills, infrastructure needs and the potential value of new bodies to bridge the gap between research genetics databases and health service clinical decision support systems.

The European Federation of Pharmaceutical Industries and Associations (EFPIA) is investigating suitable infrastructures as part of the €2 Billion European Union IMI (Innovative Medicines Initiative; see http://www.imi.europa.eu/) public-private partnership. The eTRIKS (European Translational Information and Knowledge Management Services) project has been initiated to support multiple IMI translational projects. A UK based example is the MRC/ABPI collaboration in Rheumatoid Arthritis (RA-MAP). RA-MAP exemplifies the typical challenges associated with translational research needing to bring together all existing appropriate and accessible anonymised UK RA patient clinical, genomic and environmental data into a single database to allow in depth analysis for factors related to spontaneous and therapy induced clinical remission. The data are currently held in many data sources within pharmaceutical, academic and medical establishments in a wide range of formats, with various data standards and with varying associated patient consent.

An example of what a UK infrastructure might look like is provided by the Acropolis system developed for the UK National infrastructure for translational medicine and collaboration in Cancer

(see http://www.idbs.com/data-management-news/press-release/idbs-leads-stratified-medicine-consortium-to-build-uk-wide-cancer-research-and-collaboration-platform/). Funded by the Technology Strategy Board and a consortium of Academic, Industrial, NHS and Charitable institutions, this system provides the integrated access to data, with appropriate safeguards and access rights, that enables collaboration between a diverse set of partners.

Research linking electronic health records enables the use of more effective treatments, improvements in drug safety, assessment of risks to public health and identification of the causes of diseases and disability at a speed and scale not previously possible.

It is vital that the UK research community is in a strong position to maximise the research potential offered by linking electronic NHS records with other forms of routinely collected data and research datasets e.g. UK Biobank. The MRC led a mapping exercise (MRC) on behalf of the UK research funders and the Association of British Pharmaceutical Industries (ABPI) in 2010 to review the existing UK capability and examine the requirements to support a sustainable research base in the future. The report highlighted the need to build capability and capacity in health informatics research and for further methodological development in the use of electronic records for research, including complex data linkage.

To prime research in this area the MRC has brought together a consortium of ten government and charity health research funders to fund a joint initiative to establish Centres of Excellence in research linking electronic health-related data. The Centres will create UK capability in utilising electronic health-related data for research by providing advice and expertise to the support the wider research community, encouraging interdisciplinary collaboration and offering new training and career development opportunities. The Centres will also play an important role in engaging with the public to promote better understanding of the benefits of e-health records research and will provide an interface for collaborations with industry and the NHS.

3. E-Health

E-health records research is an area in which the UK is well placed to develop a world lead. Initiatives led by the health departments in England, Scotland and Wales will provide infrastructure that will enable access to a range of federated NHS datasets in a way that has not been possible until now for both academic and industrial research.

Stewardship and use of electronic health care records (also called Real World (RW) Data) is moreover of critical importance for post-marketing surveillance/Pharmacovigilance as well as patient care. In 2011 the Association of the British Pharmaceutical Industry published a White Paper on this subject (http://www.abpi.org.uk/our-work/library/industry/Pages/Vision-for-Real-World-Data.aspx) highlighting the need for change in the current regulatory environment, to maximise the opportunities for RW data collection and analysis. The report calls for the encouragement of NHS-industry partnerships, and investment in the skills base required for research based on electronic health records. New pan-European pharmacovigilance legislation also contains a new obligation for companies to perform post-authorisation efficacy studies in certain circumstances. Sources of good quality RW data will therefore be essential for companies to meet regulatory expectations.

The Research Capability Programme (RCP) in England has launched the Clinical Practice Research Datalink (CPRD) service. CPRD - and its equivalents in Wales and Scotland, the Secure Anonymised Information Linkage (SAIL) system and the Scottish Health Information System (SHIS) –

are the three UK national programmes charged with building infrastructure that will enable the linkage of a range of electronic health (E-health) records for research purposes. The RCP will provide research infrastructure, taking account of strict information governance requirements that respect patient confidentiality, but it will not provide funding for the research, training purposes or additional linkage of non-NHS datasets. This remains the role of public funders, working alongside and in collaboration with, the national programmes in a concerted effort to maximise the potential and value of the new infrastructure for research, patient safety and public health.

The Department of Health through its existing Biomedical Research Centres and their host NHS Trusts at; Imperial, Kings, Guys & St Thomas', UCL, Oxford, Cambridge and SLaM is developing plans to deliver electronic linkage of patient records within the Trusts to support researchers, plus forging linkages into CPRD at a national level. This combination of resources will create a unique UK position for researchers.

For industry, an ability to interlink the existing NHS single access system for data mining and identification of patients for clinical studies (with the right safeguards) would provide a resource that would not only enable clinical trials to be conducted more effectively and efficiently in the UK, but also assist in managing and monitoring utility/treatment outcomes of medicines once approved (perhaps through adaptive licensing and rigorous follow up as part of a managed access system for data collection and clinical utilisation) – thereby feeding back into the discovery/development loop. This would make the UK both a fast adopter but also a rigorous data generator and thus enabling many aspects of industry objectives, patient need and government policy.

Whilst CPRD does link to some Scottish and Welsh GP data most UK data sharing is done on a case by case basis. For both industry and academe data from the UK population as a whole is important; **we recommend that:**

- **a system is developed to provide a single portal for research studies in the UK**
- **NHS Trusts prioritise the provision of dedicated funds for E-health records systems (Trust not NIHR responsibility)**

A mapping exercise in 2010 led by MRC, on behalf of the UK research funders(1) and the Association of British Pharmaceutical Industries (ABPI), to review the UK capability in e-health records research and determine the requirements to support a sustainable e-health informatics research base in the future recommended that the following areas be addressed:

- the skills shortage of people to carry out the complex linkage and analyses required in health informatics research.
- There is an absence of career structure in enabling roles such as data managers, software engineers, informaticians and data analysts.
- Methodological research in complex data linkage requires further development.
- There are no clear interfaces between researchers and industry, policy makers or the NHS and there is no ready means for sharing best practice.

To begin to address these capacity and capability issues highlighted in the report, a consortium of ten UK government and charity research funders have launched a £19m partnership to establish centres of excellence in research linking electronic health data (http://www.mrc.ac.uk/E-healthCentresCall). The Centres will undertake cutting edge research linking e-health records with other forms of research and routinely collected data, with the aim of catalysing the building of a vibrant e-health informatics research capability in the UK.

Whilst the investments outlined above are timely and welcome, there remain challenges to be addressed and action is required to :

- Enhance the completeness, system linkage and interoperability, standardisation and quality/robustness of data captured in databases
- Furthermore, the UK needs to continue to participate and have influence in international initiatives (such as the IMI Electronic Health Records for Clinical Research project; see http://www.ehr4cr.eu/) to expand development and access to a wider set of electronic healthcare records held by other countries.

Reccomendations:

- a system is developed to provide a single portal for research studies in the UK – a single access point for both academe and industry which could build on one or more of the existing national structures
- NHS Trusts prioritise the provision of dedicated funds for E-health records systems (Trust not NIHR responsibility)
- NIHR, and devolved adminstartions further develop investments in BRCs, BRUs to generate a rich UK wide data set
- Further investment is made in both skills and method development to enable the complex linkage and analyses required in health informatics research
- a career structure is developed for enabling roles such as data managers, software engineers, informaticians and data analysts
- further investment is made in methodological research in complex data linkage.

4. Chemistry

Chemistry is an essential discipline in the pharmaceutical and agrochemical industries. It is responsible for generating the majority of intellectual property on which the business models depend. This importance has not been necessarily reflected in the relative investments in e-infrastructure and from an industrial viewpoint this is something that needs to be addressed.

4.1 Bioactive molecule design

Perhaps the most familiar application of computational chemistry in the life sciences is that of structure-based discovery, whereby computational methods are used to predict the interaction of molecules (small synthetic molecules or biologicals as possible drugs or agrochemicals) to proteins. Although industry has outsourced a significant proportion of synthetic chemistry roles to China, India and Eastern Europe, the Design of molecules (with the associated generation of Intellectual Property) has remained in house. High level computational simulations are not used routinely, due to the inaccuracies in the science that can currently be accessed and the fact that the run time for the calculations does not match the cycle time for chemical synthesis and screening. Recent advances in high performance computing have enabled more accurate simulations (both molecular dynamics and quantum chemistry calculations) to be performed and it is expected that such methods will become routine in the next 5 years. Already a UK biotech company has demonstrated the use of a commercial cloud computing provider to access > 70,000 compute nodes processing ~0.5 Terabytes of chemical structure data. Such calculations have the potential to shorten drug discovery timelines from ~2 years to 1 year or shorter and provide competitive advantages for those

companies that are able to exploit the technology. Large companies are likely to collaborate with commercial providers and/or niche technology companies/academics to access such methods, whilst SMEs are more likely to evolve from work at the cutting edge to generate novel technology which will provide a differentiated competitive position. In both cases, access to knowledgeable researchers, and potential employees, in the UK will be essential.

4.2 Predictive Toxicology

Safety is a significant reason for attrition in pharmaceutical product development, with up to half of all potential medicines failing to reach the market because of unpredicted toxicology in animals or humans. The industry is engaged in several pre-competitive EU projects in order to develop new computational models to estimate molecular toxicity, through the pooling of data and collaboration with academia or SMEs. The EU IMI eTox project, OpenTox initiative and EPAA (European Partnership for Alternative Approaches to Animal Testing) are all examples of such pre-competitive work. Several UK SMEs, for example Lhasa Ltd., are core members of these projects alongside UK academics, Astra Zeneca, GlaxoSmithKline and Syngenta. Industry contributions are in-kind (people and data) amounting to many millions of pounds.

4.3 Formulation & Food Science

There are emerging applications for simulation that extend beyond the atomistic level into tissue, organ or materials modelling, generically known as multi-scale simulations. For example, in the pharmaceutical industry there is a challenge to predict the flow of drug substance in manufacturing, which involves prediction of a crystal form, the properties of that crystal and then the flow of a material composed of such crystals. Similar challenges arise in Food sciences, where Soft Matter modelling is used to predict the texture and material properties of products. Such modelling requires a coupling of deep expertise and substantial computational power. There is a limited supply of this skill set in the UK and although the Research Councils have invested through the European CECAM consortium (http://www.cecam.org/), there has been little or no adoption of the methods by the Life Sciences industry.

4.4 Data, publication & integration

The number of available chemical structures and associated data is rapidly increasing, as the public domain (particularly in the USA, but now Europe too) has invested in high throughput screening with associated collections of compounds. In the next 5-10 years it can be predicted that most of the data accessed by a pharma company researcher will not have been generated by their own company. Increasingly, the volumes of data available will mean that companies will not download and integrate the public data into their own systems, rather they will be wishing to search the data via secure access mechanisms and to integrate the results of their searches into local data analysis systems. The ~€10 million IMI OpenPhacts project (http://www.openphacts.org/) was proposed by EFPIA members with a view to developing an infrastructure to support a distributed data publication and analysis model.

The number of potential chemical structures that could be synthesised as possible drug molecules exceeds 10^{60}. An initiative to enumerate all discrete chemical structures up to 13 atoms (Blum *et al*. 2011) has yielded a database (www.gdb.unibe.ch) of size 2.6Gb. To extend this to 27 heavy atoms (which covers drug-like molecules) would require Exabyte storage and associated computational

resources (algorithms and raw computing speed) in order to search the chemical structures. This is an example where a shared resource could benefit all companies, large and small.

Most chemical reactions are still published in traditional journals, the papers manually extracted and curated to create searchable chemical databases. There exist informatics technologies, invented in the UK, which could be developed to revolutionise the publication of chemical journals. As an example of the opportunities for the UK, The Royal Society of Chemistry has already shown leadership through its project Prospect and is well placed to lead this change, should the necessary computing environment and informatics skills be assembled. The Biochemical Society has a similar initiative (Utopia docs; http://getutopia.com/). Other commercial offerings, including coverage of the patent literature, also exist.

As more and more data are generated and published, integration of the data is difficult, if not impossible, without data standards. The UK is a significant player in this field, with the presence of the European Bioinformatics Institute now extending to cover chemistry and toxicology data. Industry has also invested via the Pistoia Alliance (http://pistoiaalliance.org/), and cross-industry group which funds the development of standards in scientific data publication and software development.

4.5 Informatics

Managing, integrating, manipulating and analysing large data sets requires strong informatics skills. Informatics is a discipline which combines scientific, computer science and statistical knowledge. In the life sciences Bioinformatics and Chemoinformatics are both traditional strengths for the UK, although Chemoinformatics skills in particular are in short supply. In the future, industry will continue to require scientists with these skills, and will need to add researchers fluent in “big data” analytics whereby the scale of data requires advanced algorithms, software environments and high performance computing beyond that which is traditionally encountered in the life sciences. To complement the specialists, making software more accessible to the broader scientific community, through well designed and supported interfaces, will enable laboratory-based researchers to ask more searching questions of their data and design better experiments.

4.6 Sustainable Software development

Although the UK is a prominent player, through the EBI, in life sciences data management, it is much less strong in the development of scientific software. Most commercial and academic software used by the life sciences industry originated and is further developed in the USA or mainland Europe. This puts UK students at a disadvantage in that it is difficult to develop the deep scientific understanding required to write and extend such software, and inhibits the sustainable development of academic groups which are globally competitive. For industry it means that it can be difficult to access (for consultancy or employment) cutting edge expertise in the UK.

As an example of sustainable investment, the GROMACS project (www.gromacs.org) based at the University of Stockholm and elsewhere has produced one of the most widely used simulations code for molecular dynamics. It has taken more than 20 years of continuous effort, 500+ person years of effort and ~$27 million in funding. Some of this funding was for collaborators working in the USA and Germany. Although it is probably too late for the UK to compete with established projects like GROMACS, there is an opportunity to fund a new generation of software, particularly in bio- and

chemo- informatics and data analysis/visualisation. These skills are widely sought after in industry, which would also use the software developed in academia if it were properly maintained and supported.

4.7 UK opportunities in Chemistry

The UK has strengths that should be exploited

1. There remains a high quality pharmaceutical sector, both large companies and biotech, with strong computational functions willing to collaborate with academia.
2. The UK has made significant investments in High Performance Computing which has resulted in pockets of significant capability and expertise.
3. There are several outstanding computational chemistry and chemoinformatics research groups in UK universities with good links to industry.
4. The Royal Society of Chemistry is an international leader in chemical publications.

In order to fulfil the potential of UK scientists, the UK should create at least one, preferably more, Centres of Excellence for Molecular Design in the UK. These will bring together a critical mass of researchers to pioneer new design methods and develop sustainable software for use in academia and industry. This software would then be available and supported across the UK e-infrastructure.

5. Systems Biology

In contrast to Engineering and Physical Sciences, the more widespread recognition of the value of having a model of the system of interest is a more recent occurrence in Biology. Systems Biology is defined as the quantitative study of the interactions of the components of biological systems at different levels (e.g. cell, tissue, organ, whole body and even as far as entire patient populations; for the purposes of this report, the application of Systems Biology relates to cellular process, with higher-level human or animal applications covered in the Translational Medicine section). It uses extensively quantitative and modelling approaches and has a strong data integration component. It relies heavily on cycles between theoretical and experimental work for continuous improvements of the models and is inter-disciplinary. Potential applications are very diverse and include for example: better understanding of drug-target interactions and mechanisms of action, design, engineering of bio-products (including via synthetic biology) and development of new plant crops for a sustainable agriculture in both food and non-food sectors.

Modelling is at the core of systems biotechnology, for straightforward combinatorial reasons (Kell, 2012). If one needs to make 3 or 4 changes in a network of 1000 enzyme to increase the rate of production of a biological product substantially, the numbers of ways of doing this is ca $1.66.10^8$ or 4.10^{10}, respectively, numbers far too great at present to explore exhaustively experimentally. However, computational analysis will soon find the relevant small number of enzymes that it is necessary to modify, and it is then a simple matter to produce and test the organism. This is the strategy of choice for industrial biotechnology (e.g. Park *et al*., 2007), but requires the availability of both the models and the modelling skills.

Systems biology is using large volume of quantitative data such as genomics and other 'omics information (e.g. transcriptomics, proteomics, metabolomics), kinetics information for enzymes, and

interactions information between individual components of systems such as proteins that can be more qualitative. These diverse and high-volume data must be integrated, analyzed and modelled using a high-power computational infrastructure (hardware) and a diverse set of analytical tools (software). Models have to be reviewed and improved constantly with additional information from experimental studies to generate new hypotheses and interpretation for additional cycles of validation. Therefore e-infrastructure is an essential component of the systems biology scientific community.

In the UK there is already a strong academic community in systems biology that was fuelled, for example, with the funding (£45M) of 6 large dedicated systems biology centres by BBSRC and ESPRC in 2005 and 2006 (Imperial College London, University of Manchester, Newcastle University, University of Edinburgh, University of Oxford and University of Nottingham). Interestingly at least two of these centres (Edinburgh and Imperial College) have merged systems biology and synthetic biology. The MRC has also highlighted the use of systems approaches in medical research or "Systems Medicine" as an area for funding opportunities and the European Bioinformatics Institute (EBI) in Cambridge has groups involved in Systems Biology. Systems Biology approaches have already led to important applications on diverse industries such as health and pharmaceuticals, plant breeding, bio-fuels and the chemicals products of industrial biotechnology.

The systems biology community has been effective in developing data standards such as SBML, CellML and BioPAX. From an industry perspective having an e-Infrastructure in the UK to store and manage the data used in the context of systems biology will be very beneficial. A significant fraction of the information used in systems biology is likely to be pre-competitive and in the case of private data it should be possible to develop a security model for the industry to have controlled access and therefore maintain a competitive advantage. This infrastructure should also capture models that will be improved continuously with the virtuous cycles between computational modelling and additional experimental studies. These models can then become truly validated. Advantages of a UK wide infrastructure will also include enforcement of agreed data standards, integration with public domain tools for analysis and workflows used in systems biology. It will also enable greater interoperability between databases and tools and provide a software support mechanism. Currently too much software is developed at specific institutions. The infrastructure would also be an essential shared framework for UK scientists from academia and the industry to work together and exchange knowledge. Finally it could help to provide further education in the use of tools and models for systems biology.

A UK e-Infrastructure around Systems Biology might usefully be connected with broader infrastructure initiatives such as Elixir at the European level.

A UK e-Infrastucture for systems biology used by both academia and industry will support the development of a diverse and complementary range of skills spanning from IT to experimental biology and including mathematical biology. Since biology is moving towards systems approaches, these cross-disciplinary skills are important to the industry. Above all, we need people that are very comfortable to work across specialities and understand the specific needs and challenges of the industry.

6. Synthetic Biochemistry

The potential diversity of proteins is vast, offering an enormous array of potential functions and uses ranging from new biopharmaceuticals, antibiotics, diagnostics, industrial biocatalysts (enzymes for healthcare, home products, chemicals and food sectors), new materials and fibres. We essentially lack the knowledge base to design proteins for specific functions, which is a major barrier to the exploitation of the full diversity of functions and applications. The potential of Synthetic Biochemistry has been recognized by the Technology Strategy Board, which has made it one of the priority areas of investment.

Much of science and technology consists of the search for desirable solutions, whether abstract or realised, from an enormously larger set of possible candidates. The design, selection and/or improvement of biomacromolecules such as proteins represents a particularly clear example (Kell, 2012) . This is because natural molecular evolution is accompanied by changes in protein primary sequence that (leaving aside any chaperones) can then fold up to form higher order structures with improved function or activity. However, as well as observing the products of natural evolution, we can now make DNA encoding any protein sequence. The question thus arises as to what sequences one should make for particular purposes, and on what basis one might decide.

Straightforwardly, for a protein that might contain just the 20 main natural amino acids, with a sequence length of N amino acids, the number of possible sequences is 20^N. For N = 100 (a rather small protein) the number 20^{100} (~$1.3.10^{130}$) is already far greater than the number of atoms in the known universe. Even a library with the mass of the Earth itself— $5.98.10^{27}$ g—would comprise at most $3.3 . 10^{47}$ different sequences, or a miniscule fraction of such diversity. An array of all 4^{30} (~10^{18}) 30mers of nucleic acids made as 5 μm spots would require 29 km^2! Those proteins that have been detected as sequenced entities selected by natural evolution (leave alone the much smaller number with known tertiary structure) represent an essentially infinitesimal fraction of those. Although not all sequences could be functional, the number of known proteins is thus a tiny fraction of those that might be made (e.g. by synthetic biology) for purposes of Industrial Biotechnology.

In mechanical or electrical engineering, the means by which we make an artifact with desirable (useful) properties involves a knowledgebase of the properties and interactions of its components or modules that may then be 'bolted together', tested and modeled *in silico*, and then tested experimentally. It is not yet possible to do this for a protein because we do not yet have the necessary knowledgebase or predictive methods. The international competition recognizes this, as evidenced by the recent announcement of the Novo Nordisk Foundation Centre for Protein Research (http://www.cpr.ku.dk/). However, the UK is home to many world-leading databases for protein informatics, e.g. CATH (http://www.cathdb.info/), SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) and Pfam (http://www.sanger.ac.uk/resources/databases/pfam.html), as well as major informatics centres such as the EBI (www.ebi.ac.uk) and TGAC (http://www.tgac.ac.uk/). If we are to get to the stage where we can 'design' a protein's function on the basis of a known sequence or tertiary structure (whether measured experimentally or inferred computationally), we need to invest in the necessary underpinning technology. The present state of the art is that we can to some degree design (*in silico*) rudimentary binding agents but not at all those effecting catalysis.

Since (i) the sequences of DNA that encode a particular protein are entirely known from the sequence of the desired protein alone, and (ii) we can make any DNA sequence from

oligonucleotides bolted together via synthetic biology (even a whole bacterium), the problem is mainly one of protein informatics (and the necessary assays of function). Such experimental procedures are improving, with higher throughput systems which in some cases can select for the required trait and evolve proteins without human interventions. It is not clear how this area will mature, with different computational strategies required depending on experimental developments.

The *ab initio* design approach requires a knowledgebase and computational tools that allow us, for instance, to

- Determine secondary and tertiary (and quaternary) structures from primary amino acid sequences. (This is also known as the 'protein folding problem'.)
- Relate (sub)sequences and (sub)structures effectively and *quantitatively* to function/activity
- Allow us to visualize, align and compare the many millions of protein variants that we might make (rather than the very few that Nature happens to have selected, and that have been the focus of previous research).

In contrast, an evolutionary approach will require empirical (statistical) modelling algorithms, experimental design and visualisation tools to be fully integrated with the experimental systems.

Both approaches will require advances in algorithms, potentially significant computing power but most importantly the application of the necessary intellectual skills to the selection, dissection and solution of these problems for all (potential) proteins, not just those provided by Nature.

An industrial application: Green Chemistry

As with all other industries, the pharmaceutical industry needs to address the issues of sustainable development. The manufacture of drug substance requires significant investment in synthetic routes, to maximise yields, minimise impurities, costs and environmental impact, and Green Chemistry is seen as a promising approach. To illustrate the importance to industry, GlaxoSmithKline has recently invested £12 million in a Carbon Neutral chemistry laboratory at Nottingham University. Biocatalysis (the use of enzymes to catalyse selective chemical reactions in water at ambient temperature) is of particular interest. Enzymes are able to catalyse a wide variety of chemical reactions, and importantly, their use meets 7 of the 12 Principles of Green Chemistry defined (http://www.epa.gov/sciencematters/june2011/principles.htm) by the US Environmental Protection Agency. There are now examples of drugs using biocatalysis in their manufacture, for example a major ingredient of the diabetes medicine Januvia and a key component of the widely prescribed statin Lipitor are manufactured this way. However, it is not trivial to develop a biocatalytic system. Naturally occurring enzymes are often very selective and specific in the chemical substrates that they modify, and often the yield or efficiency of the natural enzyme does not provide a cost effective manufacturing process. Therefore, it is common to modify enzymes to produce isoforms with different specificity, yield or efficiency. This design process is often referred to as Directed Evolution, and there are two common approaches. The first is a screening approach: libraries of proteins are created using random gene mutations, and these are screened for the desired activity. The size of these libraries is typically 1,000-100,000. The best scoring mutants are kept and used to seed further random mutations, and the process is repeated a number of times. The second approach involves use of protein crystallography. The atomic structure of the enzyme is determined, and used alongside computational approaches to predict residues in the enzyme (normally around the active

site) that should effect the required change. These techniques are immature: with current methodology, a researcher might expect to discover one highly productive mutant per year. Improvements, to reduce the duration of this discovery phase from years to months, will require investment in e-infrastructure: hardware to enable longer and better simulations, statistical methods for data analysis/experimental design and scientists that understand how to best integrate these approaches into their research.

7. Knowledge Extraction & Content Management

The digital revolution sweeping through biomedical science is now well documented. Constant improvements in experimental technologies are generating data at an unprecedented rate (see http://www.ebi.ac.uk/Information/Brochures/pdf/EMBL-EBI%20Annual%20Report%202011.pdf), leading to a bottleneck in our ability to analyse and make sense of it. At the same time, an ever increasing pool of textual information is just a mouse-click away, in the form of published literature, patents, conference proceedings, competitor intelligence, blogs, tweets and many other sources of scientific discourse. Indeed, with one new scientific paper every 30 seconds scientists are already overwhelmed, before one brings in the other formal sources and the seemingly infinite scientific commentary through social media.

The ability to interrogate, analyse and formulate new hypotheses based on the body of existing science is the essence of experimental science. In recent times, the growing body of electronic information sources have made it possible to do this without performing a single physical experiment. So called “hidden knowledge” within the literature can help find new uses for old drugs such as thalidomide (Weeber *et al* 2003) and explain why costly, failed industry projects were doomed from the start (Korstanje 2003). Yet, most biomedical scientists are still trying to stay on top of the data deluge using information systems and approaches conceived decades ago. To stay competitive in the digital biology era, UK science needs to ensure it is able to gain maximum value from the resources that are available.

For an industry scientist, the challenge is clear. In addition to performing and documenting their experiments, administration, supervisory and other duties, they need to keep on top of the latest developments in their field. They can’t read everything, so how should they search for the subset upon which they should focus? Current practice is far from optimal as:

- Many searches are performed on <300 word abstracts, rather than the full content of the article. This is less than 10% of a paper’s contents.
- Different systems search differently. The scientist cannot be sure their search is thorough or indeed correct as they move from website to website.
- Most searches provide no estimation of concepts such as “relevancy” or “novelty”. They simply show that a document contains a particular keyword.

To overcome this, we need automated ways to extract existing and discover new knowledge from the growing corpus of research literature. Text mining is used increasingly to support knowledge discovery and hypothesis generation, and to help scientists navigate and drill down through the mass of biomedical literature. Its primary role is to extract explicit semantic metadata such as the following: instances of biological entities; attributes of these; instances of relations that they enter into and the nature of such relations (is a bioprocess significantly affected, positively or negatively); an author’s nuanced attitude to the relations described (is s/he *suggesting*, *criticizing*, *confirming*, …), etc. Text mining enables scientists to collect, curate, interpret and discover knowledge required

The value and benefits of text mining have been extensively discussed in a recent survey (http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx). In the UK, the National Centre for Text Mining (www.nactem.ac.uk) offers a number of biomedical services, tools and resources, and provides text mining services to enrich the functionalities of UKPubMed Central (http://ukpmc.ac.uk/) in collaboration with the EBI.

The current situation is one of extremely heterogeneous and unconnected commercial and public information silos, leading to a chaotic and confusing overall information landscape for the average scientist. While we wholeheartedly support the Hargreaves proposals regarding copyright law and text mining (http://www.ipo.gov.uk/ipreview-finalreport.pdf) and note the IPO's consultation on proposals to change the UK's copyright system (http://www.ipo.gov.uk/copyright-summaryofresponses-pdf), this is not where we see the bottleneck. Rather, substantially more effort is required to provide mechanisms to search deeply into multiple content repositories, using consistent simple language. The Pistoia alliance (a non-profit, life-science industry body), in partnership with major publishers has demonstrated the feasibility of such an independent "knowledge brokering service" (see http://www.pistoiaalliance.org/blog/tag/sesl/). The publication of the Finch Group report on expanding access to research publications (http://www.researchinfonet.org/wp-content/uploads/2012/06/Finch-Group-report-FINAL-VERSION.pdf), provides evidence of growing belief that publicly-funded science should be more accessible such that results can more easily drive innovation and growth.

At present the UK funds a lot of basic research into data and text mining systems, websites that provide access to information and specialist views across information resources. Yet, we provide much fewer resources to help scientists understand the information once they have it. Possible solutions are illustrated by concepts such as the "Biological Expression Language" (see http://selventa.com/technology/bel-framework), developed by a US biotechnology company, which show how existing information can be exploited for commercial purposes. Similarly, the Neuroscience Information Framework (see http://www.neuinfo.org/), another mainly US initiative, is an impressive effort to simplify navigation of what is an incredibly complex area. Finally, the US National Research Council recently published a report entitled "Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease" (see: http://www.ncbi.nlm.nih.gov/books/NBK91503/).

A framework to support interoperability and comparison/evaluation of text mining components has been developed by the National Centre for Text Mining, which has been adopted by, among others internationally, a major Europe-wide initiative to facilitate sharing of resources and tools (http://www.meta-net.eu/meta-share). There is thus expertise and capability in the UK to handle the interoperability issue, and the National Centre has also demonstrated ability to handle processing at scale in relation to UKPMC for life scientists. There is thus a strong foundation on which to build in relation to development of e-infrastructure for text mining in support of the life sciences. There are issues to be addressed regarding access to subscription-only journals, and any

outcome may be dependent on changes in legislation, but there is also much that can be done with open access literature in the meantime. Nevertheless, appropriate levels of investment would be needed to provide the UK with an e-infrastructure involving text mining at 'collection of collections' scale with concomitant support for applications that enable scientists to pull together the results of text mining with other big data to engage easily in knowledge discovery. Also, due attention would have to be paid to training scientists in new modes of discovery in such an environment and also to expanding the numbers of trained text mining specialists who would be required to underpin such a vision.

With the combination of Life Science industry, publishers and research base, the UK is well placed to take the lead in producing a next generation environment for information discovery and exploitation. To take advantage of the opportunity requires funding, but equally it requires leadership, co-ordination and influence on both producers and consumers as to the benefit of this investment.

8. Training & Role of the Data Scientist

"Science originates in the synergy of data, computation and human expertise" - Professor Christine Borgman, UCLA

For many years the amount of data available in life sciences has been growing rapidly, a rise driven by the constant improvements in genomic technology available to the biological research community. The first references to the "genomic data deluge" date back to at least 1993, and the deluge has grown every year since - with no sign of any slowdown in the rate of increase. For example a single organisation, the Beijing Genome Institute, has recently announced a 3 million genome project in which it will sequence a million plant and animal genomes, a million human genomes and a million micro-ecosystem genomes.

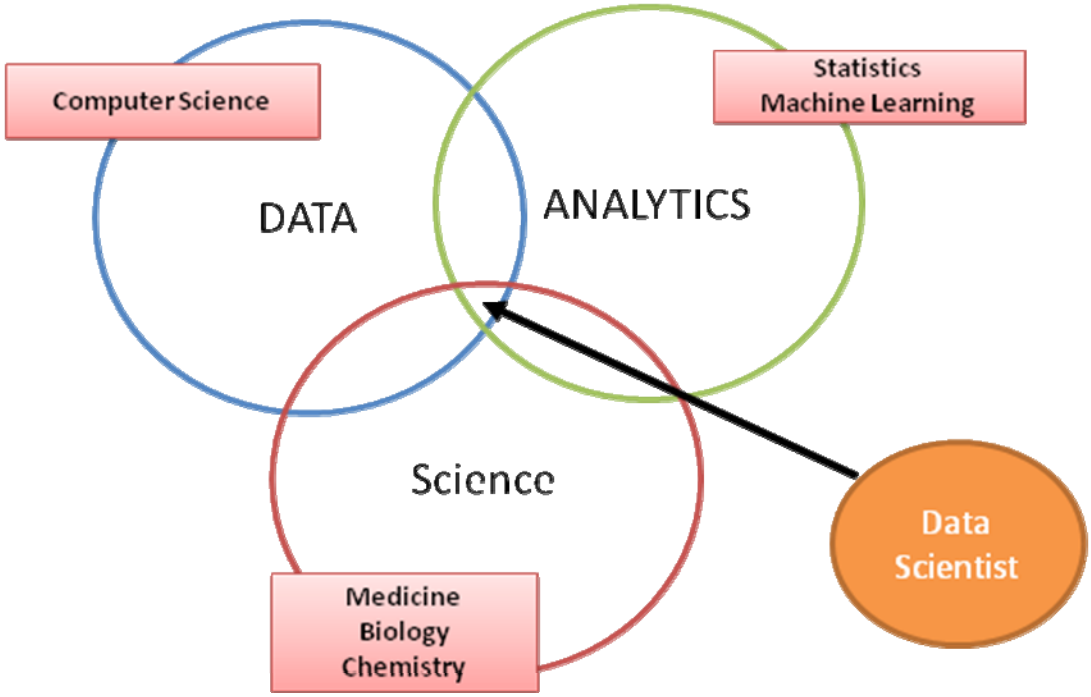
The basic data of genome sequences is only a small part of the information available to scientists. Significant data resources have been complied capturing information on many aspects of cell and organismal function ranging from gene expression to epigenetics, and from indexed literature to metabolic pathways. Biologists now have access to one of the largest, richest and most diverse data set in all of science. It is not just in biology that big and complex data sets are becoming the norm. For example, the report to Government on genomic technology from the Human Genomics Strategy Group (chaired by Professor Sir John Bell), highlighted the need to develop scientists to help manage and interpret genomics data within the NHS.

Making sense of the large and complex data sets now available requires skills that have not been traditional in life sciences, those of the "data scientist". The concept of the "data scientist" first gained currency in a number of business organisation, such as Google and LinkedIn, which collect large amounts of data from their users and are looking to see how those data can be converted into business opportunities. The skill sets required of a data scientist cover both data handling skills and data analysis methodologies. No analysis is meaningful if the data set on which it is built is full of errors and inconsistencies – so a data scientist needs the skills to collate and curate complex data and understand deeply where the data come from and their reliability. Once the data have been collected a data scientist needs the programming and machine learning skills to infer interesting knowledge from it – if necessary through the development of novel analysis techniques. One

definition of a "data scientist" is a person in which *"Columbus meets Columbo " – Monica Rogati, LinkedIn* – someone who combines the vision of the starry-eyed explorer with the scepticism of the detective. It should be recognised that such immensely skilled individuals are going to be rare beasts, and it may be better to dissect further the definition of data scientist. The person working in Google (Columbus) to digest information for others may not be the same person who is also going to be the scientific detective (Columbo). But working together and being able to communicate with each other via some shared knowledge, skills and infrastructure, the two will identify solutions to problems which would be inaccessible to either in isolation.

A recent EU report ("Riding the wave: How Europe can gain from the rising tide of scientific data"; http://www.grdi2020.eu/Pages/Unlock.aspx) has highlighted the importance of building capacity in data science to the future of science in Europe. In addition to the skills in data and analytics highlighted by industry, the report also emphasised how important it was that data scientists understood the community of science – for example the ways in which issues of trust and communication impact on the practice of science. This is an area that has been important throughout the history of science.

There are therefore three clear aspects required of a data scientist in the life sciences: skills in data, skills in analytics and an understanding of how the community of science operates.



Some of the skills required are covered in existing training schemes – for example current bioinformatics training is very good in supporting the skills for data capture and curation in biology and the Sheffield University Chemoinformatics courses provides excellent entry-level skills. Some of the systems biology training is good at helping develop new analytics to sit over existing biological data. These are areas in which the UK has traditionally taken an internationally significant role in shaping the appropriate educational programs, for example the pioneering cross-council initiatives in bioinformatics from the early 1990s, or the more recent initiatives in systems biology. The E-Science community is developing some good insights into the way in which the community of

science is organised. However, we do not have specific training programmes that encompass the integrated range of skills needed for data scientists combining strong computer science, bio/chemo-informatics, machine learning, semantics, text mining and statistics, together with an understanding from E-Science of the functioning of scientific society.

The creation of a new cadre of data scientists will therefore require an extension and expansion of existing interdisciplinary training provision. There are a number of routes in which this could be done within the UK system which build on some proven models. One obvious mechanism is through a Doctoral Training Centre mechanism, which could provide a 4-year PhD in which a significant portion of the first year is devoted to formal taught material. This mechanism has successfully developed a cadre of trained systems biologists for the UK and could readily be adapted for the production of data scientists.

At the moment there a very few specific courses that are focussed on developing data scientists – a few masters courses have started to appear in the US (for example a Masters in Analytics from North Carolina State University). This is therefore an area in which the UK could build from its expertise in bioinformatics, systems biology, chemoinformatics, E-Science and computer science education to build some strong and internationally competitive advanced level courses. This would clearly map onto the objectives set out within the European report on making best use of scientific data and e-infrastructure, and if the report to the EU commission is to be believed, will be supporting an area of rapid growth within Europe.

9. Plant Science

The plant science industry seeks to meet global challenges related to population growth, as the world population is expected to reach 9 Billion by 2050. Food production must correspondingly be increased, but in an environmentally sustainable way. The land area available for agriculture is likely to decrease rather than increase. Further challenges are posed by demographic changes in which economic development is correlated with increased meat consumption, which require higher land and water resources than current diets. In addition climate change may affect crop based agriculture in unexpected ways, further threatening the food supply. In the UK, Syngenta has the largest R&D presence and is at the forefront of efforts to develop new solutions to these challenges. In the public sector, the UK has made substantial contributions to the sequencing of a great many crop plants, most recently those of two tomato cultivars, and is leading in e.g. the 5-fold coverage of the wheat genome also announced.

The generation, integration, management and interpretation of research data is key to the overall efficiency of plant science R&D efforts. It is for these reasons that e-infrastructure (or e-science) contributes in a very direct way to the ability of developed nations, such as the UK, to respond to global challenges and to mitigate the economic and societal issues that flow from these.

9.1 Hardware / Software infrastructure.

The software and hardware infrastructure needed to manage life science data adequately is itself an area of significant challenge. The rapid development of instrumentation such as mass spectrometry and second/third generation sequencing is leading to exponential growth in the volume of data resources, in the areas of metabolomics, proteomics and (particularly) genomics. Computer

networks tend not to grow at such an accelerated rate, are highly cost constrained, and in many areas are no longer adequate for the transfer of large life science data sets. These hardware challenges arrive at a time when industrial informatics are static or shrinking. There is therefore a noticeable shift towards pre-competitive collaboration in many life science companies.

These informatics challenges affect plant sciences, in the same way that they affect many pharmaceutical groups. The challenge may be greater in plant sciences in that many important crop genomes are as yet not fully sequenced. In some instances plant genomes are larger and more complex than mammalian genomes, with higher levels of repeats regions, or variable transposable elements that make genome assembly more difficult. Plants may have variability in the number of sets of chromosomes (ploidy) that are present in the cells of a given plant variety. In addition, plant science companies must understand the biological information not only from crop plants, but from pest species such as weeds, fungi and insects. Molecular data from crop and pest species are often less well characterised than those in mammalian systems, further adding to the challenges of understanding gene function, constructing biological networks and interpreting experimental data in terms of detailed molecular entities.

A key need in this domain (and others) is to enable the development of controlled vocabularies and relationships (ontologies) that describe underlying raw experimental data. This is optimally performed in a pre-competitve manner, with collaboration between academia and industrial partners from many sectors.

High performance computing resources, drawing on large scale data storage, are needed in many areas of computation-based R&D efforts. These include genome assembly, gene sequence and function comparison and prediction. In addition small molecule capabilities are needed that provide virtual screening and docking of small molecules to proteins, large scale chemical structure comparisons, and high accuracy (quantum mechanical) calculations of chemical structure, electron distribution and molecular property (e.g. NMR, Raman spectroscopy) calculation.

9.2 Related infrastructure efforts.

In Europe the ELIXIR initiative has started the process of funding and constructing a sustainable infrastructure for life science data resources. Any future high performance e-infrastructure effort should coordinate carefully with ELIXIR. There is no dedicated and coherent and federated plant (and microbial?) science e-infrastructure effort in the UK or Europe, to the best of the author's knowledge. In the US, the iPlant collaborative is seeking to construct a cyber-infrastructure and tools to address grand challenges in plant science.

The UK does have some considerable strengths in plant e-science, however, such as the integration led by Garnet, data resources of the Nottingham *Arabidopsis* Stock Centre and others, various programs (such as ONDEX) at Rothamsted Research, world leading modeling of plant development at the John Innes Centre, emerging large-scale phenomics and e-infrastructure at IBERS, and an extensive investment at TGAC on the Norwich Research Park, providing excellent resources on which to build.

10. Animal Science

The UK has particular strengths in the Animal Sciences. All the Faculties of Veterinary Medicine in our Universities offer teaching and research facilities. The Roslin Institute at the University of Edinburgh and the Institute of Animal Health at Pirbright, represent the UK's main centres of excellence in Animal Sciences as applied to poultry, pigs, cattle and other livestock. The Roslin Institute provides post graduate teaching and research in all livestock and companion animal species, in particular poultry, pigs, cattle and fish. Roslin also hosts the National Avian Research Facility partly funded by the BBSRC to provide the research community access to avian chicken lines and research facilities. Roslin has a wide range of skills and research interests, including quantitative genetics, genomics and genetics, bioinformatics, and infection and immunity. The IAH has a special interest in viral infections and immunity, in a wide range of species including poultry, pigs and cattle. Both institutes collaborate extensively together, and with the animal breeding/animal health sectors. TGAC (Norwich) a BBSRC sponsored institute provides generic skills and resources for the UK plant, animal and microbial research communities in genomics and bioinformatics. Finally, the EBI/Sanger institutes funded by the EC, Wellcome Trust and other bodies (including BBSRC) provide a wide range of genome databases for the research communities throughout the world. The Roslin Institute has collaborated with EBI/Sanger for the past 8 years on chicken, pig, cattle and other livestock genomes with funding from the BBSRC.

Collaboration between the academic community and the animal breeding /animal health sectors occurs through a number of routes (direct, LINK, IPA, etc). Interactions are facilitated through the Knowledge Transfer Network (KTN), which hold workshops and conferences, and industry clubs.

Draft reference genome sequences have been established for several farmed and companion animal species, including chicken, cattle, pig, horse, turkey and dog. Annotated genomes for these species are available in Ensembl and other genome browsers. Sequencing of the genomes of several other farmed animal species is well advanced (e.g. sheep and duck) or in progress (e.g. salmon, buffalo, goat, deer and quail).

The scientific challenge in livestock and poultry genetics, as applied to breeding and health can be reduced to the simple equation "Phenotype = Genotype x Environment". The goal of genetics and genomics of livestock is to predict the phenotype as accurately as possible from our knowledge of the genetic variation in their genomes and the environmental context in which they exist. In this way we can manipulate the phenotype by either selection of specific genotypes or modify the environment (or both).

New sequencing technologies have facilitated detailed characterisation of genetic variation in farmed animal genomes. Genome sequence data from farmed animal species over and above the reference genome sequence are already in the public and private domains, and are expected to increase at a rapid rate. Currently these include sequences derived from multiple lines and breeds of chicken, cattle and pigs. These genome sequences can be mined for a wide range of genomic variants, including SNPs, small insertion-deletions, copy number variants and larger structural variants. Knowledge of their sequence context (coding, conserved etc) can be used to predict the functional consequences of such genetic variants on the phenotype of the animal.

Similarly, sequencing has been used to characterise transcriptomes of many organisms under many different conditions using the RNAseq approach. This method is expected to complement and possibly replace array based methods for genome-wide gene expression analysis. RNAseq can capture the transcription of the entire genome, both the coding and non-coding regions. The latter are poorly covered in traditional expression arrays. RNAseq is also an approach that can be used with any species without any array resources; the genome sequence is the only requirement. Such approaches can define host responses to infection by pathogens or effects of the environment, such as day length on host physiology. Since RNAseq is based on sequence data, it can also define genetic variants showing altered patterns of gene expression. Again, knowledge of their sequence context can help to predict functional consequences such as altered splicing or other post-transcriptional effects.

Whilst the raw sequence data necessary to establish a genome sequence assembly can be acquired quickly and at modest cost, annotating the assembly and making it accessible to the user community in a useful manner remains a significant undertaking. The proliferation of other sequenced based data (genetic variant, transcriptional etc) also requires integration of multiple and heterogeneous data types, which is a huge challenge, unique to biology.

The Ensembl genome browser, associated annotation tools and genome database infrastructure have been shown to be robust and effective means for making genomic information available to a wide range of users. (How well it will scale to the analysis of a great many genomes in parallel is yet to be determined.) Ensembl provides high quality genome sequence annotation for a range of vertebrate species, including not only humans, mice and other key model organisms, but also our farmed and companion animal species. The annotations include evidence-based gene models, predicted gene models, whole genome multi-species alignments, annotations of orthologous genes, annotations of microarray probes sequences from array platforms, single nucleotide polymorphisms, RNA-seq, Chip-seq data and links between Ensembl resources and many other reference biological databases such as UniProt, pfam and others. Many users explore the genome of their species of interest via the Ensembl genome browser. Data mining is enabled with Ensembl BioMart. Many users also download the data directly and incorporate Ensembl gene annotations and other data into their own analysis routines. Power users can interrogate the underlying database(s) directly using the Ensembl API.

Equally important to our knowledge of genome variation and content is information on phenotypes on poultry, pig, cattle and other animal populations. As outlined above, we have the genomics technology to characterise the genomes of individuals in greater detail; this needs to be raised to the population level. However to exploit this genetic variation we need to link it with phenotypic variation. Current strategies (e.g. based on genome selection) require the analysis of 1000s and even 10,000s of animals to recognise animals of high breeding merit. However this raises the problem of creating and maintaining appropriate databases to manage and provide access to these data. There is also great concern in the private sector on security and confidential data. So these databases need to capture and integrate large datasets from multiple sources, private and public, very heterogeneous in nature.

In summary, the future needs of Animal Sciences as applied to livestock and companion animals will include the following:

(i) Data management is already an issue. Datasets are huge and getting larger, and more heterogeneous. Data integration of public and private data sources is required not in data warehouses but distributed as hub and spokes. Genetic and phenotypic databases are needed on 1000s and 10,000s of animals to facilitate the analysis of complex traits. These resources are required to train and create genome predictors for a wide range of traits. These predictors can then be used in other populations.

(ii) Data analysis of these integrated resources requires new software and data models. This requires training, and easy access to flexible software.

(iii) Data modelling is required to exploit these information resources to predict consequences of genetic selection on individuals and populations, on their performance, their health, their impact on the environment, their economic impact etc.

(iv) Training is crucial, we need a new breed of researcher for both industry and academia, able to handle and mine large data sets, and create and use models to predict outcomes and impact.

11. Imaging

Medical imaging adds value to patient care, to academic research, and to industry through, for example, clinical trials. A variety of imaging modalities and associated specialised acquisitions provide a broad range of tools for clinicians and researchers. These allow for the examination of structural and functional information that are not available by superficial observation. The information gained can relate to disease, injury, treatment strategy, anatomy, physiology, drug or medical device characteristics. There are roughly 1000 imaging departments in the UK performing such scanning. In the NHS, there has been a 3-fold increase in the number of diagnostic scans performed by CT and MRI in the last decade (see http://www.dh.gov.uk/en/Publicationsandstatistics/Statistics/Performancedataandstatistics%20/HospitalActivityStatistics/DH_077487). In the pharmaceutical industry, image data is increasingly used for methodology development, bio-distribution, safety monitoring, and efficacy assessment.

It is estimated that medical imaging information storage constitutes one-third of global storage demand. Advances in technology appear to be driving an increase in the volume of data collected. Such advancements include higher resolution imaging, increased specificity, increased sensitivity, decreased scan time, as well as the ability to perform dynamic scanning where multiple datasets are acquired as a function of time. (For those modalities where there is no patient exposure to ionizing radiation, there is also a tendency to collect more data than the absolute minimum required.) Equally, any improvements in storage and networking technology lead to the increased use of applications that produce large data sets. That is, the availability of better technology allows for the performance of better radiology and research.

The key infrastructure needed is primarily network bandwidth, and secondly high performance storage capacity, but also data standards, data compression, strong encryption and security.

Centralised High Performance Computing resources are of less importance, as image analysis lends itself well to distributed computing. Developments such as IHE's (Integrating the Healthcare Enterprise; see http://www.ihe.net/) cross-enterprise document sharing (XDS) and the imaging-specific extension (XDS-I) allow the storage of data at or near the acquiring site with the existence of data being published in central registry. Authorized partners are then allowed to retrieve data. A key advantage to this approach is that data does not need to be store in multiple locations, and only data required need to cross the network. Examples of a large scale central repository approach include the proposed Biobank (http://www.ukbiobank.ac.uk/) in the UK and the Maine Health Information Exchange (http://www.hinfonet.org/news-events/news/maine-hie-pilot-nation%E2%80%99s-first-statewide-medical-image-archive) in the US. It is anticipated that research partnerships would benefit being able to interrogate multiple shared repositories for data of interest permitting novel, large scale data mining.

The UK has leading imaging experts in both industry and academia. The presence of these experts on EU projects, such as the Innovative Medicines Initiative's Quantitative Imaging in Cancer: Connecting Cellular Processes with Therapy (http://www.eortc.be/services/doc/EUprojects/QuickConcept.html), highlights their leading role in the field, and willingness for collaboration. Investment in infrastructure is required in order to maintain this position, however, investment in people in this specialized area is also critical. For example, the DICOM standard and technology (Digital Imaging and Communications in Medicine; see http://medical.nema.org/) allows for much heterogeneity provided a basic set of guidelines is followed, creating a significant challenge for data management, and the need for highly specialized data managers and programmers. From an industry perspective, resourcing such expertise is surprisingly difficult given the number of graduates in this area. (Although this section is mainly about medical imaging, we note the increasing interest and investment in imaging-based phenomics in plants, such as the recently announced infrastructure at IBERS (http://www.phenomics.org.uk/). In a similar vein, detailed imaging of the response of cells to small molecules is widely practised as 'high-content screening'.)

12. Summary and Recommendations

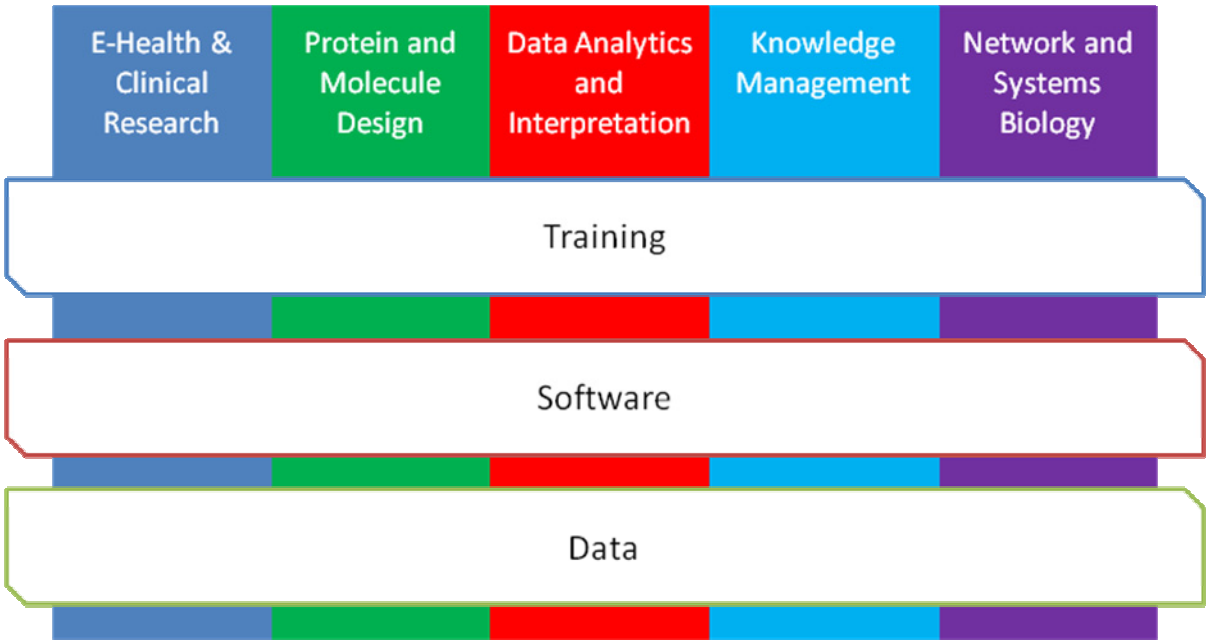
UK companies will gain advantage if they are amongst the first to exploit the opportunities afforded by the coming developments in e-health, analytics and simulation. For SMEs it can be hard to understand what e-science can do for them, and which academic groups might help them. Although larger companies mostly rely on commercial software and internal hardware, changes to the current business model, with more collaborative R&D conducted with Universities and Hospitals, highlights the needs for shared data standards, improved security and common software platforms. In particular, Translational Medicine/e-health is a priority area for focus.

Access to hardware is not the primary concern of the Life Sciences industry, which is not to say that the current investments could not provide value, rather that without equivalent investment in other components of the e-infrastructure eco system (software, people and skills), it will be difficult for the industry to exploit the enhanced hardware and network capability provided.

This report has highlighted the diverse and pressing needs of the Life Sciences industry. Given the desire for greater application of e-science in industry and the significant investments already made by government, what is preventing the required uptake? We recognise several barriers to progress:

1. For e-health applications, there is not yet a national infrastructure to ensure ethical, secure access to appropriate patient data and to enable collaboration between academics, industry and NHS clinicians.
2. The UK Pharmaceutical and Agrochemicals sectors have been significantly downsized in recent years. The commercial environment for both sectors is challenging and this limits the potential for industrial investment in long term infrastructure.
3. There is no overarching strategic vision from industry, HEFCs or the research councils which can be used to plan investment.
4. Computational science has been added to lab-based science departments in Universities, which have traditionally not integrated well. As such, there are no "e-science" departments offering undergraduate (or even postgraduate) degrees.
5. Central investment in high performance computing has focussed on limited hardware and software applications. Life Sciences have a set of diverse needs and this has resulted in a fragmentation of the UK infrastructure.
6. Previous investment has been primarily in physical capital. The Life Science sector places greater emphasis on Human Capital – scientists who can manage and interpret the coming deluge of data – and the generation of software for the many informatics tasks. (The annual summaries of database and software resources in the journal Nucleic Acids Research run to ca 1000 of each.)
7. Lack of career paths for scientific software developers and database curators in academia.
8. Pressure in industry mean that the traditional routes for academic collaboration – CASE awards or post-doctoral funding – are either too long term or too expensive. More immediate access to academic expertise for shorter periods is required.

As this report has illustrated, what is termed the Life Sciences sector is actually a highly diverse collection of disciplines and applications. It is helpful to categorise our recommendations not by sector but by activity, and to consider the e-infrastructure requirements for each activity as summarised in the graphic below:



There is a consistent recommendation which is common across all the activities: **we must invest in the UK skills base through better training.**

There is a need for increased skills in health informatics, systems biology, Protein and Molecule design, librarianship, legal compliance, information management, text analytics, semantics & semantic technology, social media analytics and visualisation.

We recommend that

- Multiple specialist Masters level conversion courses are established, appropriate to the sectors
- Shorter courses are made available for Continuing Professional Development of industrial scientists.
- A career structure is developed in academia for enabling roles such as data managers, software engineers, informaticians and data analysts
- E- Science modules should be incorporated into all Life Science undergraduate degree courses

For each activity there are also specific recommendations.

Health and Clinical Research

Software

- A priority is the development of a system to provide a single access portal for research studies in the UK.
- The UK should develop an infrastructure that enables academics, hospitals and industry to collaborate in an environment that realizes the required ethical standards, privacy policies and data access rights. At the very least, standardised dictionaries, ontologies and software interfaces should be agreed on and consistently applied.
- Further investment is needed in methodological research for complex data linkage.

Data

- NHS Trusts need to prioritise the provision of dedicated funds for E-health records systems with high quality data (this should be a responsibility of the Trust). The National Institute for Health Research, and devolved administrations, should further develop investments in

Biomedical Research Centres and Biomedical Research Units to generate a rich UK wide data set.
- Deliver current investments in data services such as CPRD and SHIP.
- Targeted training/engagement of healthcare professionals is needed to encourage use of electronic health records.

Network and systems biology

Software

- A priority is investment in the development of the tools of network and systems biology that are usable by working biologists without a computer science background.

Data.

- Support the Open Data agenda together with suitable interoperable data models that will allow the federated delivery of data describing biochemical systems and that then allow integration with our literature knowledge to provide semantically annotated models of biological systems.

Protein and Molecule Design

Training

- Investment in hardware and computer networks needs to be suitably complemented by investment in people with life sciences domain knowledge. All e-infrastructure should have people dedicated to facilitate training of scientists and enable access to the appropriate high performance computing resources for their scientific problem (the "On Ramp").
- Consider a Knowledge Transfer Network dedicated to e-Science.

Software

- The e-infrastructure should host a mixture of commercial and open source software. This will reduce duplication of effort in academia and at the same time increase robustness and supportability for industry. Create at least one, preferably more, Centres of Excellence for Molecular Design in the UK, which will produce a critical mass of researchers to pioneer new design methods and develop sustainable software for use in academia and industry. This software would be available and supported across the UK e-infrastructure.

Data

- The e-infrastructure must be designed and implemented such that proprietary data can be routed and stored securely. Without complete confidence in the fidelity of the infrastructure it will not be possible to attract industrial use.

Data Analytics and Interpretation

Software

- In order to be competitive the UK must invest in the development of novel algorithms. This will drive improvements in data analytics, machine learning, advanced statistics and visualisation.

Data

- While data science assumes (and requires) the availability of the relevant data, it will be important that the relevant data are in fact available (the Open Data agenda) together with the necessary metadata set out according to recognised standards of interoperability. Assistance for developing, and where necessary imposing, such standards will be required.

Knowledge Management

Software & Data

- The UK should provide core Knowledge Management services over UK generated content and beyond. Immediate opportunities should centre around the concept of 'The UK Book of Science' covering all UK funded reports, scientific papers, grants, patents, conferences etc. Initial services could include 1) UK Key Opinion Leader analysis enabling users to find out who is working on what innovation enabling UK based collaborations or 2) Derived Knowledge Services allowing users to rapidly navigate automatically generated knowledge summaries for any scientific concept(s) to support hypothesis generation.
  Such KM services would require a maintained core compute infrastructure, federation and integration technologies, standards agreement and implementation and copyright & content access agreement.

E-science is at the core of the Bioscience knowledgebase. Only those nations that recognise and exploit this will prosper.

13. Works Cited

Ananiadou, S. *et al*. (2010). Event extraction for systems biology by text mining the literature. *Trends in Biotechnology* 28, 381-390.

Blum L. C., van Deursen, R. & Reymond, J.L. (2011) Visualisation and subsets of the chemical universe database GDB-13 for virtual screening. J Comput Aided Mol Des 25, 637-647.

Harland L., Forster M. (2012). Response from IMI Openphacts project to the RCUK vision & roadmap.

Hull, D. Pettifer, S. & Kell, D. B. (2008). Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Comput Biol* , e1000204.

Kell, D. B. (2012). Scientific discovery as a combinatorial optimisation problem: how best to navigate the landscape of possible experiments? *Bioessays* 34236-244.

MRC. (n.d.). www.mrc.ac.uk/e-health.

Park, J.H. *et al*. (2007): Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and *in silico* gene knockout simulation. Proc Natl Acad Sci U S A 104, 7797-7802.

Stratton M.R., Campbell, P.J. & Futreal, P.A. (2009). The cancer genome. *Nature* 458, 719-724.

Weeber M. *et al* (2003): Generating Hypotheses by Discovering Implicit Associations in the Literature: A Case Report of a Search for New Potential Therapeutic Uses for Thalidomide. J Am Med Inform Assoc. 10, 252–259

Korstanje C. (2003): Integrated assessment of preclinical data: shifting high attrition rates to earlier phase drug development. Curr Opin Investig Drugs. 4(5),519-21.

14. List of Contributors, Consultants and Reviewers

*Darren Green	GlaxoSmithKline
*Douglas Kell	BBSRC
*Ian Dix	AstraZeneca
*Robert Glen	University of Cambridge
*Peter Coveney	University College London
*Anne-Marie Coriat	MRC
*Lesley Thompson	EPSRC
Sophia Ananiadou	University of Manchester
Andy Brass	University of Manchester
Klaus Bengst	Astra Zeneca
David Burt	Roslin Institute
Andy Ellis	Biocats
Jonathan Essex	University of Southampton
Wendy Filsell	Unilever
Paul Finn	Inhibox
Chris Foley	GlaxoSmithKline
Mark Forster	Syngenta
Keith Godfrey	University of Southampton
Lee Harland	Independent Consultant
Vincent Hughes	Fujitsu UK
Janette Jones	Unilever
Peter Knight	Department of Health
Amanda Lane	Unilever
Chris Larminie	GlaxoSmithKline
Louise Leong	ABPI
Nick Lynch	Astra Zeneca
Ian Mitchell	Fujitsu UK
Chris Molloy	IDBS
Paul Mortensen	Astex Pharmaceuticals
Christine Orengo	University College London
Chris Page	GlaxoSmithKline
Anthony Rowe	Johnson & Johnson
Philippe Sanseau	GlaxoSmithKline
Willie Taylor	NIMR
Janet Thornton	EMBL-EBI
Jon Wrennall	Fujitsu UK

*Members of ELC