Challenges of ultra large scale integration of biomedical computing systems

R.H.J. Begent
NCRI Informatics Unit, London, UK
A. Finkelstein
University College London, UK
P. Kerr
NCRI Informatics Unit, London UK
F. Reddington
NCRI Informatics Unit, London UK
J.M. Brady
Oxford University, UK
D. Gavaghan
Oxford University, UK
H. Parkinson
NCRI Informatics Unit/EBI, UK
J.M. Wilkinson
NCRI Informatics Unit, London UK

Abstract

The NCRI Informatics Initiative is overseeing the implementation of an informatics framework for the UK cancer research community. The framework advocates an integrated multidisciplinary method of working between scientific and medical communities. Key to this process is community adoption of high quality acquisition, storage, sharing and integration of diverse data elements to improve knowledge of the causes, prevention and treatment of cancer. The integration of the complex data and meta-data used by these multiple communities is a significant challenge and there are technical, resource-based and sociological issues to be addressed. In this paper we review progress aimed at establishing the framework and outline key challenges in ultra large scale integration of biomedical computing systems.

1. Introduction

The application of computational tools in cancer research and cancer care has become a vital and rapidly developing activity, responding to the large amount and great diversity of data being generated. The field of bioinformatics has primarily developed to address this for high throughput or data-rich areas of research such as genomics or proteomics. However, large datasets are also produced in cell biology, physiology, pathology, imaging, therapeutics, clinical trials and epidemiology. With current information systems, we are not in a position to make maximal use of existing, or future, data sets generated by cancer research.

The National Cancer Research Institute (NCRI) has created a vision of an internationally compatible informatics platform in the UK that facilitates access to, and integrated analysis of, data generated from cancer research, across the spectrum from genomic data to data generated from clinical trials. A strategic framework was developed as a first step towards making the vision a reality. If successful in cancer, it could be extended to other areas. Key elements of the vision are:

- The development and adoption of common data standards.
- Development of a culture of data sharing.
- Development of a website (www.cancerinformatics.org.uk) and use of an online 'Planning Matrix' as a practical tool to map information resources and bring stakeholders with common interests together.
- Establishment of strategic partnerships with key organisations such as the US National Cancer Institute (NCI) and the European Bioinformatics Institute (EBI).
- A number of demonstrator projects to show what can be achieved.
- Development of common policies for data management amongst NCRI partners.
- Reviews of current infrastructure and training.

- Communication and consultation with the cancer community.

This will be supported by:

- A High Level Steering Committee to provide strategic guidance.
- An expert Task Force, composed of domain experts, to oversee implementation.
- An NCRI Informatics Coordination Unit to facilitate and influence.

Whilst the vision is undoubtedly a worthy one, a number of key challenges must be addressed if the Initiative is to succeed in using informatics to maximise the impact of cancer research.

2. The current situation

Historically a reductionist approach to scientific research has prevailed. We are now moving towards a time whereby an integrative approach is needed to fully exploit the knowledge that can be gained from the data that already exists – so called systems biology and systems medicine. This type of environment will require both a culture of data sharing and an academic rewards system that acknowledges large multi-disciplinary collaborations rather than personal advancement. The NCRI has chosen to address this problem in two ways. Firstly, the NCRI are working with the funding organisations to develop a joint data sharing policy (www.cancerinformatics.org.uk/documents.htm#datasharing). This policy advocates that publicly funded research data should be openly available to the maximum extent possible unless consent or ethical approval have not been gained, the confidentiality of study participants can not be safeguarded, or there are intellectual property issues that are of concern to a potential or existing commercial partner. The NCRI data sharing policy has now been formally adopted as policy by Cancer Research UK, the world's leading charity dedicated to research on the causes, treatment and prevention of cancer. Encouraging researchers to think about data sharing strategies at the time of applying for grants will promote a shift towards a data sharing culture where appropriate.

Secondly, the NCRI has convened a diverse Task Force, sanctioned by the High Level Steering Committee, to identify challenges and areas for collaboration. One of the key challenges faced by the Task Force is that whilst there is a desire for the construction of a platform from the user community, clearly defined use cases for the platform do not yet exist. This is partly because the scientific and clinical questions, which the platform will enable the community to address, are dynamic and flexible.

The NCRI have funded a demonstrator project, with a clearly defined use case, that illustrates the challenges of the systems approach.

2.1 Pathophysiology and imaging demonstrator

The pathophysiology and imaging proposal is a data driven proposal. This demonstrator comes from radiological, microscopy, clinical trials, computer science and integrative biology communities. It develops a framework that re-uses and adapts systems developed for imaging of breast cancer and applies them in rectal cancer, integrating magnetic resonance imaging (MRI) information with macroscopic images, microscopy and data from a clinical trial. After testing and validation in this application, the framework will be made available in an open source format to the research community so that it can be used with comparable datasets or adapted for use with other imaging modalities or types of clinical data. In this way it will be a

critical tool to enable implementation of the NCRI partners commitment to data sharing, re-use and integration. In day-to-day terms delivery of the above might, for example, enable the following use case:

A query is made to the system for patients meeting certain criteria. Images are retrieved for the relevant set of patients in real-time from the individual databases (possibly behind local firewalls). A single patient is chosen and all relevant images displayed. Services are invoked to undertake a 3D reconstruction of the macroscopic resected rectum. Initially this will be from the photographs but will be extended to also build in the virtual histology allowing correlation of in vivo pathology with the cellular pathology. A diagnostic feature of tumour close to the margin or suspected high risk features such as venous invasion or peritoneal invasion is highlighted in one image and a further service is invoked to identify the same feature (and highlight) in other images and within the 3D reconstruction. It would be advantageous if pre-treatment scans and post treatment scans could be fused with differences automatically highlighted. The data needs to be seen at Leeds, the Royal Marsden, Basingstoke (Pelican) and Oxford. Patient identifiers will be removed and substituted with trial code number held only by the trials office.

3. Challenge 1

Construction of large scale integrative platforms can not be done centrally, and the NCRI Informatics Initiative will not build the platform, instead it will enable the community to build the platform. The role of the NCRI Informatics Initiative is to enable this process and devising a work plan that allows that to happen synergistically is non trivial.

3.1 Response

The NCRI Informatics Initiative Coordination Unit brings together specialists from different disciplines, in a Task Force, to act in a targeted way to identify and address problems common to projects in cancer biomedicine. This Task Force is unique in the breadth of domain experts it encompasses and has shown that a multidisciplinary approach to implementation of such a large-scale Initiative is both feasible and productive. By establishing a network of resources and expertise we can tackle problems such as data integration in clinical trials and functional genomics collectively, as illustrated in a recent NCRI workshop (http://www.cancerinformatics.org.uk/workshops.htm). This allows individuals to move beyond project-specific workplans and to refine these workplans collectively based on their respective experience. This represents an efficient use of resources and a move to an integrated multidisciplinary method of working.

4. Challenge 2

A diversity of existing bottom up individual research projects already exists in the cancer domain and the Initiative thus needs to work in this context, starting from scratch is not an option.

4.1 Response

Historically in biomedical research the development of IT infrastructure has been a data driven ‘bottom up’ approach whereby individual research projects develop specific local solutions to data storage and acquisition problems. This is evident by the ubiquity of the ‘flat file’ in bioinformatics, which was, until recently the preferred mode of data exchange. Given

the prevalence of high throughput technologies flat files and local bespoke systems are not a scaleable solution for the management, exchange and integration of large and complex datasets, such as those generated by genomics, proteomics and metabolomics. Neither are they a useful visualization tool. The NCRI Informatics Unit has surveyed the state of the art in informatics around cancer and produced a web enabled matrix:

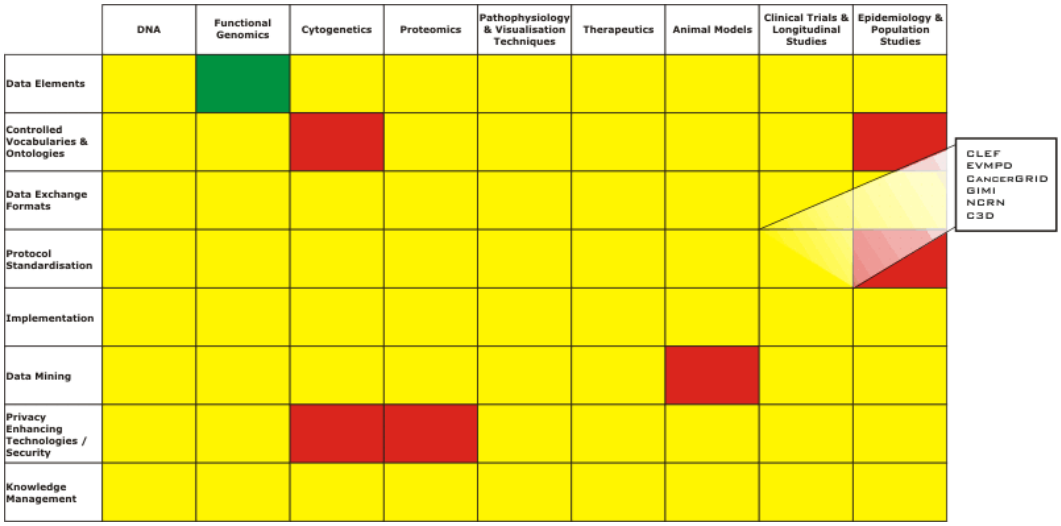

Figure 1. NCRI planning matrix
www.cancerinformatics.org.uk/planning_matrix.htm.

This matrix is available online and has been designed to display information and identify areas of commonality between projects, resources and domains involved in cancer research. However, the matrix is generic in its applicability and can be applied to any major area of biomedical research (e.g. cardiovascular disease). Along the horizontal axis of the matrix are the key communities of interest, loosely ranged left to right by dominant scale of concern. This is a diverse set of domains, but many of the informatics problems are common to many, if not all, and the use cases traverse multiple domains. Definitions of the domains can be found at (www.cancerinformatics.org.uk/matrix_definitions.htm). On the vertical axis are broad areas in which these communities have made progress on matters related to information access and sharing. The degree of progress is colour coded.

The matrix is populated by instances within cells and each instance has additional information in the form of a summary and a URI. The matrix is primarily used as a resource identification tool at present and provides a powerful quick glance summary, valuable for planning, but has limited application beyond this bundling together as it does standards, projects, specific databases etc. We have therefore used the instances in the matrix as the basis of an object modelling approach to develop a Platform Reference Model that identifies the key components of the existing infrastructure and their relationship to each other. The NCRI have funded a demonstrator project that will construct a simple metadata repository that will implement the Platform Reference Model (composed of both domain and infrastructure models). The Platform Reference Model is currently being evaluated and refined with use cases provided by the various stakeholder communities. The use cases will refine the Platform Reference Model and also help us evaluate what infrastructure is required for similar projects. By this top down process with the data providers developing use cases we can move iteratively to a full specification for an informatics platform, based on what is needed now and in the future. It will also allow us to re-use resources – skills and software

and perform high level integration across the diverse set of projects that are currently funded. The metadata repository could then be populated and used alongside the matrix to track ongoing work. As work on the platform progresses the repository could serve as the forerunner of a service and component broker.

5. Challenge 3

Emerging technologies such as the Grid and the semantic web are not yet proven in the biomedical domain or in a production or service context. This makes the architecture of a platform of this type difficult to establish.

5.1 Response

The grid community itself is active in the evaluation of the current technology and are working to develop coherent approaches to common issues, for example those related to Web Services (http://www.nesc.ac.uk/technical_papers/UKeS-2004-05.pdf) and security (http://www.nesc.ac.uk/technical_papers/UKeS-2004-04.pdf). Additionally grid projects from the first round of a UK-wide call are maturing and are making an impact on the biomedical user community. (Stevens, R. D., Robinson, A. J. and Goble, C. A. (2003). myGrid: personalised bioinformatics on the information grid. *Bioinformatics* **19 Suppl 1**: i302-4.). The Clinical e-Science Framework (CLEF) is a notable example, as it is now entering a services phase (http://www.clef-user.com/).

6. Challenge 4

The cancer domain has many diverse and dynamic standards for data elements, semantic reporting, and data formats and many of these are developing independently of each other.

6.1 Response

The Initiative is working with the relevant communities to identify, evaluate and promote the adoption of common standards within the domains of the matrix. This process is occurring within the functional genomics community, for example the Minimum Information about a Microarray Experiment (MIAME) standard has spawned an ontology and a data exchange format and is now a paradigm for functional genomics. The birth of MIAME was not hampered by legacy standards and the technology was relatively new however this is not the case for all domains, and in some cases multiple formats exist, e.g. clinical trials.
The Initiative is also working to ensure that standards between domains develop in a complementary way. This is exemplified by the current integration efforts between Microarray Gene Expression (MAGE) and Health Level 7 (HL7). MAGE facilitates the exchange of microarray information between different data systems and through an object model and mark-up language and there is currently a group within the HL7 workspace (an international healthcare standard) investigating how to integrate MAGE with the HL7 Reference Information Model. Where diverse standards exist, the Initiative is working with major National and International projects to ensure interoperability, compatibility and coherence.

7. Challenge 5

Data ownership is a somewhat confused issue in the biomedical domain. Funding organisations pay for the generation of data but the resources and sense of ownership reside

with the primary researcher rather than specialist centralised databases with data management resources. The data may be made public at the time of publication, which is often some time after data acquisition by which point quality control is problematic.

7.1 Response

A cultural change is required to encourage a large-scale collaborative working environment where individuals receive credit for sharing of resources and data. As previously discussed, encouraging researchers to think about data sharing strategies at the time of applying for grants will promote a shift towards a data sharing culture where appropriate. Furthermore, the Initiative is conducting an infrastructure review to document the current status of databases that appear in the planning matrix and will highlight areas where key resources are needed to enable data sharing.

8. Challenge 6

It is essential to the long-term strategy of the NCRI's Informatics Initiative that training mechanisms coexist with any large-scale implementation framework. As standards and infrastructures are implemented individuals must efficiently and accurately acquire the knowledge to use them. Researchers and clinicians will be required to collaborate more often and in larger groups, which will be composed of individuals willing to contribute multi-discipline solutions to problems outside their area of expertise.

8.1 Response

A review is being conducted that will identify programmes or resources that embody the principles of the Initiative. Underlying the structure of the review are two considerations. First is a focus on training specific to cancer research (in order to manage the complex and diverse training domain), second is a more general approach which aims to leverage resources not directly associated with cancer research. The Initiative seeks to establish a forum for key persons engaged in cross-discipline training to exchange ideas, identify areas of focus and keep abreast of latest developments. The training review will also inform potential demonstrator projects of existing training resources and mechanisms.

9. Conclusion

In conclusion, the work of the NCRI Informatics Initiative to date has shown the importance of community involvement in the development of an Informatics Framework for the cancer community. The development of a Task Force, data sharing policy and planning matrix provide us with vital resources upon which to build. The Task Force and Coordination Unit will deliver a report to the NCRI Board in Autumn 2005 that will describe the current status of cancer informatics in the UK, and will encompass detailed reviews regarding training and infrastructure. The report will consist of recommendations from the community about how best to take UK cancer informatics forward in a complementary fashion with other National and International Initiatives, such as the UK Clinical Research Collaboration (UKCRC) and the caBIG Initiative in the US. During the preparation of the report work will continue to further elaborate upon and validate of the Platform Reference Model, which will serve as the basis for a cancer-wide meta data repository.