BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 1. 52–63
Advance Access published on 25 March 2010

doi:10.1093/bib/bbq010

Online resources of cancer data: barriers, benefits and lessons

Emanuela Gadaleta, Nicholas R. Lemoine and Claude Chelala

Submitted: 29th January 2010; Received (in revised form): 1st March 2010

Abstract

With advances in high-throughput techniques, the volume of data generated has resulted in the creation of a plethora of resources for the cancer research community. However, a key factor in the utility, sustainability and future use of a novel resource lies in its ability to allow for data sharing and to be interoperable with major international cancer research efforts. This article will introduce some of these efforts, the interoperable cancer data-mining resources and repositories, from a user-perspective. Some of the considerations to be addressed when building interoperable, sustainable cancer resources will be discussed with case studies—hoping this will prove useful for researchers designing their own cancer databases.

Keywords: *data-mining; heterogeneity and isolation of data; interoperability; online cancer resources*

INTRODUCTION

The diversity and rapid evolution of high-throughput technologies, used to elucidate the genetic alterations associated with tumourigenesis and development of resistance to treatment, have resulted in the generation of a myriad of data values [1, 2]. A critical factor in the pace of advancement lies in the ease with which the cancer research community is able to share, locate, extract, analyse and integrate such valuable accrued data into their own research [3].

Currently, cancer data tend to be stored in independent research databases, each with proprietary data formats and each not fully compatible, nor integrative, with other resources [2, 4]. Similarly, the absence of a community-wide adoption of standard vocabularies results in the generation of heterogeneous representations of biologically similar datasets [3]. This ‘Tower of Babel’ problem is endemic in cancer (and other fields of) research and inhibits effective usage of the valuable existing data [1].

Interoperability is vital for the utility and productive use of any current or future sustainable cancer database. Interoperability refers to the capacity of multiple independent systems to be interfaced and comprises syntactic and semantic components. While syntactic interoperability ensures that there is the capacity for data exchange, through shared interfaces, semantic interoperability is the ability of a system to access and interpret the exchanged data [5].

Only interoperable resources will ensure that data generated across different organisations can be shared to maximise the impact of the underlying research and avoid duplication of effort. This would permit the design of more sophisticated portals capable of enabling efficient mining of the stored cancer data while accelerating biological interpretation and scientific discovery of new relationships among factors contributing to the complex pathogenesis of cancer.

Acknowledgement of the pressing need for an interoperable bioinformatics structure has resulted in multiple groups aiming to implement novel integrative technologies and software systems into cancer research [1]. A key challenge in any data integration solution would be its ability to address multiple cancer types and subtypes generically.

Corresponding author. Claude Chelala, Centre for Molecular Oncology and Imaging, Institute of Cancer and CR-UK Clinical Centre, Barts & The London School of Medicine (QMUL), Charterhouse Square, London EC1M 6BQ, UK. Tel: +44-20-78-823570; Fax: +44-20-78-823884; E-mail: c.chelala@qmul.ac.uk

Emanuela Gadaleta is an Assistant Researcher in Bioinformatics at Barts & The London School of Medicine (QMUL).

Nicholas R. Lemoine is Director of the Institute of Cancer at Barts & The London School of Medicine (QMUL). He also leads the Centre for Molecular Oncology & Imaging.

Claude Chelala is a Lecturer in Bioinformatics at Barts & The London School of Medicine (QMUL), and the Bioinformatics Lead at Barts Cancer Research UK Centre.

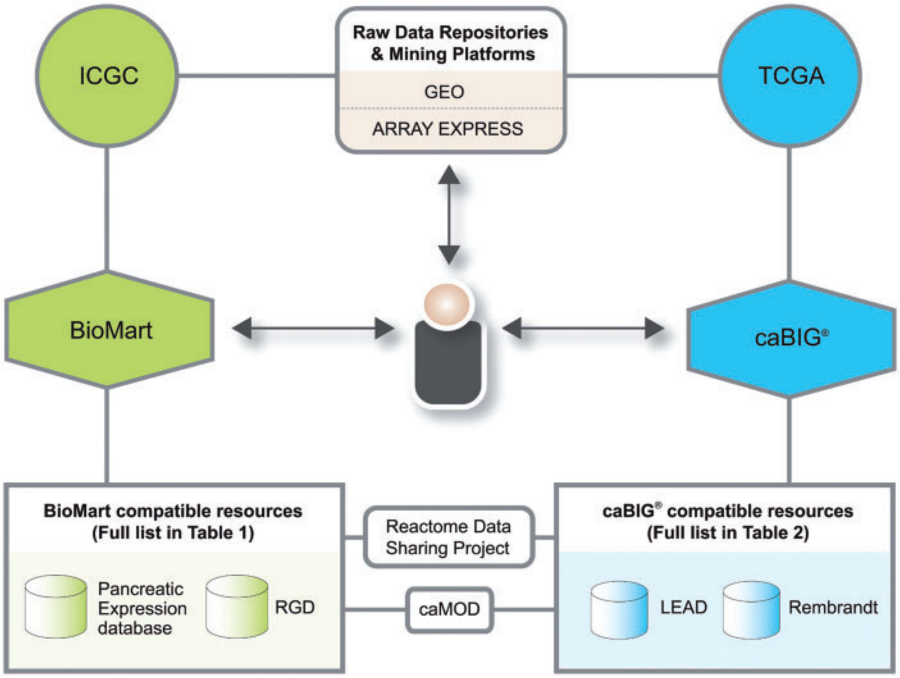
ICGC
Raw Data Repositories & Mining Platforms
GEO
ARRAY EXPRESS
TCGA
BioMart
caBIG®
BioMart compatible resources (Full list in Table 1)
Pancreatic Expression database
RGD
Reactome Data Sharing Project
caMOD
caBIG® compatible resources (Full list in Table 2)
LEAD
Rembrandt

Figure 1: A flowchart representing how the user can interact with the prominent gene expression repositories, the major international cancer research efforts with their underlying data management systems as well as independent compliant resources. This illustrates the benefits of ensuring compliancy with BioMart and/or caBIG® when designing a novel resource.

This review, undertaken from a user perspective, will discuss the major international cancer research efforts, their underlying data-mining systems as well as prominent repositories (Figure 1). We hope that the issues addressed, and case studies used, will prove useful for researchers looking to build interoperable and sustainable cancer databases.

MAJOR INTERNATIONAL CANCER RESEARCH EFFORTS

Tumourigenesis has been attributed to the amalgamation of numerous interacting events [6]. While the traditional approach to cancer research has been to study isolated components, recently there has been an evolution to a global-systems approach—for instance, integration of profile data at the genomic, transcriptomic, proteomic and metabolomic level [7].

To achieve this integrated global profiling, the barriers between research initiatives, associated with lack of compatibility, communication and coordination, must be overcome [1, 2]. To this end, international cancer research efforts, such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), are being launched and dedicated to the systematic study of alterations in a variety of human cancers. Integration and data-mining of datasets obtained from these initiatives is expected to lead to more discoveries and a better understanding of cancer development, diagnosis, treatment and prevention.

The Cancer Genome Atlas

The TCGA project (http://cancergenome.nih.gov/) [8], developed through a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), is a large-scale, US-based initiative aimed at the molecular characterisation of the spectrum of genetic alterations associated with the initiation and propagation of cancer [9, 10]. Genomic alterations reported to be associated with malignant transformation, including changes to the DNA sequence, copy number alterations, chromosomal aberrations and epigenetic modifications, will all be assessed by the TCGA [11].

Originally designed to study brain, ovarian and lung cancers, the TCGA pilot project has focused its efforts on brain cancer (glioblastoma multiforme) and ovarian cancer (serous cystadenocarcinoma). The future expansion plan of TCGA will aim to encompass and characterise more than 20 tumour types over the next 5 years.

Currently the official point of access to the TCGA is its own data portal, available from the TCGA homepage. This portal acts as a web-based platform for syntactic interoperability; that is to say, the point from which users are able to access, query and download the TCGA datasets. Additional access is provided through the Cancer Genome Workbench [12], the UCSC cancer genome browser [13] and the Cancer Genomics Data Portal, at the Memorial Sloan–Kettering Cancer Center (MSKCC) [14].

In the same manner, the Cancer Molecular Analysis portal (CMA) [15] provides access to analytical tools to mine, integrate and visualise the TCGA clinical and genomic data. This allows for the effective exploration of relationships between patient information, such as age, tumour burden and survival rates, and molecular data, such as somatic mutations and DNA methylation status. Kaplan–Meier survival plots, box plots, heatmaps and simple pathway diagrams can readily be generated to display the various dimensions of the clinical and molecular data simultaneously. Furthermore, tools such as principal component analysis and gene pattern analysis allow for in-depth interrogation of the datasets.

With the Cancer Biomedical Informatics Grid (caBIG®) forming the data-mining infrastructure of the TCGA, it was important that this initiative be designed to conform to the underlying compatibility criteria. For instance, caBIG® has been adopted to connect and coordinate the participating research centres, with all data generated being compliant with caBIG® data standards. Because it is a component of the caBIG® initiative, the CMA portal can be used to connect, collect and share information obtained from compliant systems.

TCGA data are now accessible from the services-middleware infrastructure, caGrid, ensuring greater accessibility and integration of datasets. For example, data from the TCGA and the Pathway Interaction Database (PID) will shortly be accessible from caGrid. This link allows for a combination of information obtained from the TCGA with features accessible via the PID, thereby empowering scientists with the tools required to conduct in-depth downstream analysis. There will be the potential to explore signalling pathways of interest and create interaction network maps, focusing on specific entities. The heightened connectivity between these resources will increase the potential for data dissemination, aiding the effectiveness of cancer research [16, 17].

International Cancer Genome Consortium

The ICGC (http://www.icgc.org/) [18] is a complementary, relatively newer but distinct, resource to the TCGA. Established in 2008 as a consortium of funders and research institutions worldwide, the main goal of the ICGC is to generate a comprehensive, and publicly available, catalogue of somatic genomic, transcriptomic and epigenomic aberrations in tumours obtained from 50, globally important, cancer types or subtypes. Control data, obtained from matched non-tumour tissue, will be used to differentiate between somatic aberrations and germline sequence variations. From the selected tumour types, about 500 cancers of each class will need to be sequenced in order to achieve the aim of detecting cancer genes with somatic abnormalities in at least 3% of a single cancer type. In addition, the consortium will produce catalogues of transcriptomic and epigenomic information from these tumours [19].

The goals set by this initiative will be achieved by the ICGC, coordinating and overseeing both a network of current and future initiatives, such as the Cancer Genome Project (CGP) [20]. While initiatives such as the TCGA participate with the ICGC through the sharing of data and an overlapping number of research efforts, each resource will remain a separate entity and continue to operate independently. The ICGC Data Coordinating Centre (DCC) will act as an umbrella organisation, managing the flow of complex data in the project across the participating centres and making it publicly available to the entire research community [21].

Because of the geographically distributed nature of the ICGC project, and the diversity of the data produced, the Consortium recently announced the adoption of the federated BioMart technology for its data management system architecture [21]. This will ensure that each participating repository is capable of developing individualistic data models [21–23]. Each of the participating institutions will perform comprehensive research into specific cancer types or subtypes, ensuring that they conform to ICGC DCC standardisation guidelines. The ICGC DCC will then provide a common infrastructure for the information obtained from the different centres such that there will be homogeneity in data query and retrieval across systems, with the information produced appearing to originate from a single portal—regardless of the institution from which the data were generated [21, 23]. Currently, the open access data tier will be accessible via the ICGC web portal, with the user being able to perform interactive queries and analytical tests, on the available datasets, for specimens of interest, as well as to visualise and download related data files. Expansion of the options available to this querying system is projected for the near future.

DATA-MINING SYSTEMS

In this section, BioMart and caBIG®, the underlying data management systems of the aforementioned cancer research efforts, will be discussed further. In addition, the Pancreatic Expression database, The Lymphoma Enterprise Architecture Data-system (LEAD) and the Repository of Molecular Brain Neoplasia Data (Rembrandt) will be used as proof of concept to highlight the benefits of compliance with these federated schemas. The final part of this section will introduce the recently released ONcology Information eXchange (ONIX) portal, which aims to form a connection between datasets obtained from multiple data management platforms.

BioMart

BioMart (http://www.biomart.org/) [22] is a publicly available data management platform. Initially developed as a joint collaboration between the European Bioinformatics Institute (EBI) and the Ontario Institute for Cancer Research (OICR), BioMart is a simple federated query system based on a generic framework designed for biological storage and retrieval [23, 24]. The generic nature of BioMart contributes to the robustness of this data management platform with regard to its ability to be used by researchers looking to establish open systems. Furthermore, the universal architectural similarities of the BioMart-powered resources enable users to navigate between them effectively and with ease.

All BioMart-powered resources can be accessed via the BioMart central portal (Table 1) [22]. Here, the data can be integrated into the third-party software packages Bioclipse [25], Galaxy [26], Cytoscape [27], Taverna [28], WebLab [29], Ruby API [30] and the purpose-designed Bioconductor package, biomaRt [31]—for easy interrogation within the open source R statistical environment (Figure 2) [32]. This integration enables biostatisticians to conduct both integrated queries with profiling experiments and in-depth analysis of results. BioMart-compliant resources could be configured as Distributed Annotation System (DAS) servers [33]. DAS can be used as a publicly available facility, capable of collating and exchanging biological information obtained from multiple servers, providing DAS annotations for the wider community so they can be used in other resources or browsers, such as Ensembl GeneView using GeneDAS protocol [34].

Biologists and bioinformaticians can access BioMart-powered resources through the URL-based query interface, MartView, and programmatically through web services [22, 24]. The multiple levels of access offered by BioMart (Figure 2), and all BioMart-powered systems, ensure its exposure to a scientific community encompassing a wide range of disciplines. In addition, complementarity between BioMarts ensures effective information connectivity and sharing between the resources, with any improvements to BioMart contributing indirect benefit to these systems.

The Pancreatic Expression database

Currently, as the sole cancer-based database originally designed to be BioMart-compliant, the Pancreatic Expression database (http://www.pancreasexpression.org) [35] has facilitated the integration and rapid data-mining of pancreatic cancer literature data [36, 37]. Optimal for downstream analysis, the database offers an extensive range of selections, allowing for broad and specific interrogation of the complex cancer datasets in the repository.

The Pancreatic Expression database provides information on gene expression measurements from precursor lesions and various stages of pancreatic cancer. The datasets incorporated into the database were obtained from highly relevant pancreatic cancer publications, originating from multiple international laboratories, and comprise a multitude of different platforms.

In addition to the multi-access levels permitted by all BioMarts, the provision of a Pancreatic Expression Linkout annotation by NCBI EntrezGene enhances the capacity for exposure of this open system.

By permitting researchers to mine and integrate datasets from the database into their own research, the Pancreatic Expression database serves not only to

Table 1: Publicly available Marts

Resource	Description	URL
Ensembl	Database of genomic data for vertebrates and other eukaryotes	http://www.ensembl.org
Ensembl Bacteria	Annotated bacterial genome database	http://bacteria.ensembl.org
Ensembl Metazoa	Annotated metazoic genome database	http://metazoa.ensembl.org
Ensembl Protists	Annotated protista genome database	http://protists.ensembl.org
Ensembl Plants	Annotated plant genome database	http://plants.ensembl.org
Ensembl Fungi	Annotated fungi genome database	http://fungi.ensembl.org
Phytozome	Plant genomic repository	http://www.phytozome.net
Gramene	Repository of grass genomics	http://www.gramene.org
Europhenome	Repository of raw murine phenotypic data	http://www.europhenome.org
UniProt	Repository of annotated protein sequences and functional data	http://www.ebi.ac.uk/uniprot
InterPro	Database of protein and associated data	http://www.ebi.ac.uk/interpro
HGNC	Repository of approved gene symbols	http://www.genenames.org
CyanoBase	Cyanobacteria genome database	http://genome.kazusa.or.jp/cyanobase
Wormbase	*Caenorhabditis elegans* and related nematode genome database	http://www.wormbase.org
DroSpeGe	Annotated Drosophila genome data	http://insects.eugenes.org/DroSpeGe
ArrayExpress DW	Warehouse for gene expression microarray data	http://www.ebi.ac.uk/gxa
Eurexpress	Transcriptomic database for mouse *in situ* expression information	http://www.eurexpress.org/ee
International HapMap Project	Catalogue of common genetic variants in five populations	http://hapmap.ncbi.nlm.nih.gov
Dictybase	Genomic and proteomic data for the *Dictyostelium discoideum*	http://www.dictybase.org
Rat Genome Database	Model organism database for the rat	http://rgd.mcw.edu
GermOnLine	Database of information for genes associated with sexual reproduction	http://www.germonline.org
PRIDE	Proteomic data repository	http://www.ebi.ac.uk/pride
PepSeeker	Database of peptide identification and ion data	http://www.ispider.manchester.ac.uk/pepseeker
VectorBase	Invertebrate vectors of human pathogenesis	http://biomart.vectorbase.org
HTGT	High-throughput gene targeting to produce knockout mice	http://www.sanger.ac.uk/htgt
Pancreatic Expression Database	Database for pancreatic cancer expression data-mining	http://www.pancreasexpression.org
Reactome	Curated set of core biological pathways	http://www.reactome.org
EU Rat Mart	Compilation of rat tissue expression data	http://www.ebi.ac.uk/euratools/euratmart.html
Paramecium DB	Database for the model organism *Paramecium tetraurelia*	http://paramecium.cgm.cnrs-gif.fr
International Potato Centre (CIP)	Germplasm passport and evaluation data	https://research.cip.cgiar.org
Mouse Genome Informatics	Database of integrated genetic, genomic and biological mouse data	http://biomart.informatics.jax.org/biomart/martview

aid the elucidation of the pathobiological events underlying pancreatic cancer but also contributes to the identification of diagnostic and therapeutic targets, cross-platform meta-analysis and the development of novel diagnostic tools.

The Pancreatic Expression database provides a unique tool to the cancer research community. Unlike many databases, specialised in providing single-type information, the Pancreatic Expression database is capable of retrieving and integrating both genomic and proteomic expression data, thereby resulting in a higher dimensionality of data [38]. For instance, the Pancreatic Expression database can be used to search and retrieve genes, or proteins, expressed only in pancreatic cancer, and not in chronic pancreatitis, and then query which of these were present in the urine and/or plasma proteome. Such a query would be a first step for the non-invasive discovery of pancreatic cancer biomarkers. In addition, cross-platform meta-analysis can be performed. Scientists can retrieve the sets of overlapping genes between their own results, obtained by a particular platform, and those reported in the studies stored in the database. These examples of use are well documented in the Pancreatic Expression database papers [36, 37].

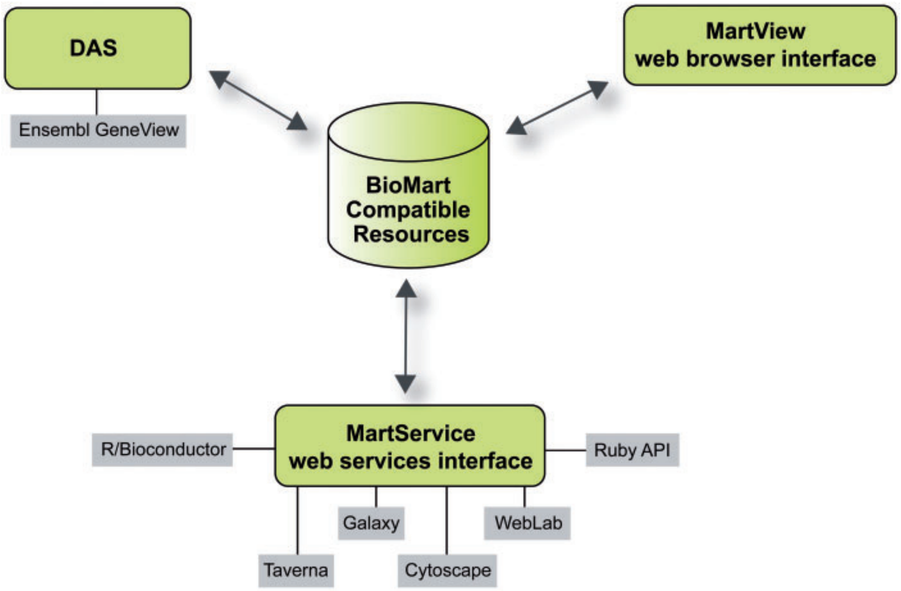

Figure 2: An illustrative example of the multiple levels of access offered by BioMart and its compliant systems.

Synergistic results could be obtained via compliance between the pancreatic expression datasets and those of other resources using BioMart technology. For example, compatibility with the ICGC DCC will allow the integration of multiple ICGC experimental datasets with the literature datasets obtained from the Pancreatic Expression database—a process that would be enhanced by the projected developments of novel BioMart interfaces, capable of facilitating the simultaneous data-mining of numerous datasets, and expansion of storage options [24].

The generic extensible BioMart infrastructure provides the Pancreatic Expression model with the architectural flexibility required for limitless expansion potential, meaning that this concept could be extended to encompass information associated with multiple cancer types.

Cancer Biomedical Informatics Grid

caBIG® (https://cabig.nci.nih.gov) [39], established by the NCI in 2004, aims to provide a common informatics platform to the cancer research community via the seamless integration of heterogeneous datasets and the provision of open-access interoperable tools (Table 2) [40, 41]. The founding network architecture of this virtual informatics platform is caGrid. Similar to BioMart, caBIG® is based on a federated approach—coordinating the information obtained from independent, compatible institutions [5].

In order to connect with caBIG®, interoperable software systems need to be created and adopted. An assessment of interoperability is conducted, and qualified, based on four maturity levels, ranging from no interoperability capability to full semantic and syntactic interoperability—classified as legacy, bronze, silver and gold, respectively. The two complementary pathways provided to achieve interoperability are to 'adopt' or 'adapt', with the selection being dependent on the needs and existing infrastructure of the institution. There is also the option of a hybrid approach, in which an institution may choose to both adopt novel caBIG® tools while adapting some of the pre-existing infrastructures to fulfil the criteria required for caBIG® compatibility.

With many tools available to connect to the caBIG® infrastructure, the requirement for a standard system between these tools is necessary for both semantic and syntactic interoperability [3, 5]. To ensure compliance and interoperability with caBIG® technology, novel resources must achieve predefined data standards.

Coherence of the caBIG® system ensures successful cross-model comparisons of equivalent data elements obtained from different models [3]. From a biologist's perspective, even with the possible connectivity difficulties associated with comprehensibility and ease of navigation, there is no question as to the worth of the information obtainable from caBIG®.

As with BioMart, the flexibility of the caBIG® infra-architecture ensures that this data management system and, by proxy, any resources based on this system, is not limited to the representation of a

Table 2: Some caBIG®-compliant resources and tools

Resource	Description	URL
Cancer Genome Anatomy Project (CGAP)	Resource of annotated genes associated with cancer evolution	http://cgap.nci.nih.gov
Cancers Genetic Markers of Susceptibility (CGEMS)	Research program aimed at identifying common genetic variants that confer to susceptibility to cancer	http://cgems.cancer.gov
caBench-to-Bedside (caB2Bcancer)	Query tool enabling researchers to search and combine data from a multitude of caGrid services	https://cabig.nci.nih.gov/tools/caB2B
caArray	Array data management system allowing for the sharing and integration of data across caBIG	https://cabig.nci.nih.gov/tools/caArray
caNanoLab	Data portal permitting access and dissemination of nanoparticle information	http://cananolab.abcc.ncifcrf.gov/caNanoLab
Cancer Genome-Wide Association Scan (caGWAS)	This tool enables researchers to mine associations between genetic alterations and disease, therapeutic response and clinical outcome.	https://cabig.nci.nih.gov/tools/caGWAS
Cancer Molecular Analysis Portal (CMAP)	Allows for the analysis of genes associated with oncogenesis, cancer profiles, clinical trials and therapies	http://cmap.nci.nih.gov
Investigation of Serial Studies to Predict Your therapeutic response with Imaging and Molecular Analysis (I-SPY Trials)	Resource aimed at identifying biomarkers and surrogate markers predictive of response to therapy in women with stage 3 breast cancer	http://tr.nci.nih.gov/iSpy
The Lymphoma Enterprise Architecture Data-system (LEAD)	Database integrating clinical and translational lymphoma data	http://umlmodelbrowser.nci.nih.gov
Cancer Models Database (caMOD)	Data management system allowing for the management and sharing of information obtained from animal models	https://cancermodels.nci.nih.gov
The National Biomedical Imaging Archive (NBIA)	US repository integrating *in vivo* cancer images with clinical and genetic data.	https://cabig.nci.nih.gov/tools/NCIA
Rembrandt	Repository for molecular brain neoplasia data	https://caintegrator.nci.nih.gov/rembrandt
The Cancer Genome Atlas (TCGA)	Data associated with the molecular characterisation of genetic alterations associated with brain, lung and ovarian cancer	http://cancergenome.nih.gov
Reactome	Curated set of core biological pathways. Development of Reactome as a data feed to caBIG	http://www.reactome.org

specific cancer but instead has the potential to be extended to encompass additional cancer types and other complex diseases. This would enable researchers to conduct complex integrated queries across all systems in the federation from a single access point. Furthermore, all caBIG®-compliant resources are interoperable with each other and will benefit from any improvements to the infrastructure, including the development of novel tools [42].

The Lymphoma Enterprise Architecture Data-system

LEAD (http://umlmodelbrowser.nci.nih.gov/umlmodelbrowser/) [43] is a database capable of integrating clinical and translational lymphoma data from a multitude of participating organisations. This system facilitates information connectivity and data sharing by unifying data obtained from these disparate institutions into a single structure. To date, LEAD is the sole lymphoma database registered using the caBIG® management system [44].

Having fulfilled the rigorous semantic and syntactic interoperability criteria, LEAD has qualified as a silver level compliant system. The data standards fulfilled include provision of Application Programming Interfaces (APIs) to access the data and tools; use of controlled vocabularies and data elements to ensure homogeneity, continuity and comparability between systems; and registered metadata covering clinical and biological lymphoma research [44].

When performing a query, LEAD can be accessed either by using the web browser interface or programmatically. The user is then able to interrogate the database for information such as tumour type, histological information, demographic details of the patient, and focus on the treatment provided—including any adverse effects and patient outcome.

Repository of Molecular Brain Neoplasia Data

Rembrandt (http://rembrandt.nci.nih.gov) [45], a password-protected, publicly available bioinformatics system, was developed in 2005 to tackle issues associated with accessibility and integration of brain tumour biomedical data [42]. This database stores both clinical and molecular data and provides a data-mining and analysis platform for queries. The Rembrandt design is based on caGrid, thereby ensuring compatibility with caBIG® and all resources compliant with this infrastructure, such as the TCGA.

Public data can be accessed either programmatically through the Rembrandt caGrid service or via the web browser, CMA and the TCGA homepage. This database provides the ability to perform queries based on gene expression, copy number analysis and clinical data, with retrieved results able to integrate information obtained from the different data domains.

A concrete example of use might be to ask which genes in the EGF signalling pathway are differentially de-regulated in astrocytoma samples profiled using Affymetrix GeneChip® Human Genome U133 Plus 2.0 platform. Application of a clinical filter, to the obtained results, restricts analysis to those samples with the specified criteria (for example age, gender, tumour grade, etc.). Rembrandt also allows users to correlate the expression profiles of the de-regulated genes with the genomic aberrations in the selected tumour samples. The Rembrandt platform permits users to collect, interrogate, import and share additional data obtained from other resources such as GenBank and Biocarta pathways.

ONcology Information eXchange

Developed in 2009 by the National Cancer Research Institute (NCRI) Informatics Initiative, ONIX (http://www.ncri-onix.org.uk) [46] is an open source portal that focuses on generating an extensive catalogue of existing cancer-related initiatives. The portal provides users with a single point of access to query datasets generated by participating cancer initiatives and, by assembling these diverse resources in a single site, increases their exposure to the cancer research community [47].

In conjunction to using the NCRI's own approach, the ONIX portal was designed to sit on a caGrid platform. Furthermore, provision of interoperable adapters to non-compliant resources allows for seamless information flow between these, otherwise isolated, entities and caBIG®-compliant resources, thus providing additional unity and homogeneity to the query retrieval process.

Queries are currently posed through text matching, although scheduled improvements aim to implement a semantic query capability [48, 49]. This tool will, initially, be designed to allow researchers to interrogate data across all caBIG®-compliant resources. However, the ultimate goal is the generation of an interoperable platform capable of the complex, federated, querying of datasets obtained from a culmination of international initiatives.

GENE EXPRESSION REPOSITORIES AND DATA-MINING PLATFORMS

The local long-term preservation and sharing of the cumulative volume of gene expression data being generated could prove problematic, placing strain on the infrastructure used by the institution. To facilitate access and conservation of microarray data, without the need to store the data locally, the raw data can be uploaded to a permanent, publicly available repository, such as the Gene Expression Omnibus (GEO) or ArrayExpress [50].

GEO and ArrayExpress require submissions to be uploaded in Minimum Information About a Microarray Experiment (MIAME)-compliant format to facilitate the subsequent integration of data to different resources, allowing for valid comparisons between datasets [51, 52]. In addition to providing the user with the raw and normalised datasets, adherence to MIAME specifications also ensures access to protocol and to experimental details. The reference for each data source is available, to trace the origin of the data. This allows for access to further clinical and pathological sample information and enables users to check if the data complies with their quality standards.

The primary data are stored and provided with an accession number, which acts as a reference for future studies. Both integration of uploaded datasets into novel studies and the application of algorithms and statistical methods by future research efforts could aid in the discovery of new, meaningful, results.

Several scientific journals now have strict data sharing guidelines and require authors to be willing to share their data with independent investigators and deposit their raw datasets in GEO or ArrayExpress as a condition of publication. Unfortunately, many authors still do not follow these sharing policies [53]. In order to help protect intellectual property interests, both GEO and ArrayExpress will keep data private for up to 1 year or until the manuscript, referring to the data, is published [54].

This section will discuss the prominent repositories, GEO and ArrayExpress. It will then highlight the power and limitations of mining the data stored in the repositories using Oncomine.

Gene Expression Omnibus

GEO (http://www.ncbi.nlm.nih.gov/geo/) [55], established as a public repository containing high-throughput data, permits the effective exploration, query and visualisation of gene expression data [52, 56]. The flexible architecture of GEO results in a versatile design capable of accommodating heterogeneous data. GEO is not cancer-specific and attempts to encompass profiles from all studies using microarray technologies. In addition to data obtained from microarray methodologies, GEO also contains data types obtained from non-array-based techniques, such as serial analysis of gene expression (SAGE) [57].

Once the data have been submitted to GEO, they are subjected to manual inspections ascertaining content integrity and syntactic validation, with these submissions being subsequently stored as platform-type, sample-type or series-type components and allocated an accession number [51].

While data analysis for the rapid identification of potentially significant information can be accomplished using auxiliary tools provided, these functions are not recommended for robust systematic analyses [56, 58]. Downloading and reanalysing raw data in a statistical environment can address this weakness in the system. For example, use of the Geoquery software package forms a bridge between Bioconductor and GEO, with the software tool directly accessing the information stored within the repository and allowing for rigorous analyses to be conducted [32, 59]. The strength of this approach lies in the eradication of formatting and parsing issues.

ArrayExpress

Array express (http://www.ebi.ac.uk/arrayexpress-as/ae/) [60] is also an international public repository for manually curated microarray data [61, 62]. As with GEO, this open source system is not limited to cancer data. The versatile design of ArrayExpress allows users to undertake data-mining of published gene expression data. This resource allows users to access the repository using the web interface or programmatically through APIs, ensuring its appeal to biologists and bioinformaticians alike.

ArrayExpress is divided into an archive of experimental data, containing all annotated microarray data supplied in published articles, and a warehouse of additionally curated subsets of data, obtained from gene expression profiles. Depending on the nature of the query, the user is able to interrogate either sub-component of the repository from the query interface. Multiple features and additional experimental details are available subsequently to the user. As with GEO, the availability of raw data means that researchers are able to download the datasets into Bioconductor for further analysis or run in-house analyses [63].

The benefits of collaboration between GEO and ArrayExpress have been recognised, with ArrayExpress now capable of importing and integrating Affymetrix and Agilent-based microarray gene expression data from GEO [61, 62]. Furthermore, the ArrayExpress infrastructure is being developed to increase the repository's capacity for dealing with the increasing volume of data generation and its ability to manage publications based on advancing platform types.

Interoperability of the gene expression warehouse will also be increased by its replacement by the enriched, BioMart-powered, Gene Expression Atlas [64]. This novel repository of meta-analysis-based summary statistics will reap the benefits associated with BioMart compliance whilst permitting the user to conduct more expansive searches and analysis from a greater volume of datasets than currently accessible via the ArrayExpress warehouse. With a larger number of statistical packages being compatible with BioMart-powered resources, statistical options will also increase significantly with the implementation of the Gene Expression Atlas [64].

Oncomine

Oncomine (http://www.oncomine.org/) [65] is a password-protected, web-based, data-mining platform of cancer microarray data, used predominantly for the upstream analysis of cancer profiles [66]. While Oncomine makes available datasets downloaded from GEO and the Stanford Microarray

Database (SMD), which subsequently are standardised, it itself is a closed system.

The collection of microarray studies, obtained from published literature, is manually reviewed, collected and prioritised. Unlike GEO and Array Express, Oncomine does not allow for submission or download of datasets. The lack of access to raw data means that researchers are unable to alter the standardisation pipeline. However, Oncomine does provide the original PubMed citations from which the original publication can be accessed. Rather restrictively, data analysis is limited to groups that have been pre-defined in the literature.

The sole site of access into the database is via the web portal. When posing a biological query, the availability of comprehensive tutorials aid the user-friendly data-mining approach provided by this database [67]. Oncomine can either be browsed in its entirety, with the option of user-defined restrictions to a search, or interrogated using gene-centric and profile-based queries. In-depth analysis by focusing on specific areas or profiles of interest, using the analysis functions provided by Oncomine, can be conducted subsequently. Alternatively, user-selected profiles can be combined using meta-analysis for use in cross-study validation.

CONCLUSIONS

With caBIG® and BioMart dominating as cancer data management platforms, the generation of a software bridge between these technologies could revolutionise cancer research and increase the utility and sustainability of these and other cancer resources. While waiting for the implementation of this bridge, researchers could design their resources to be compliant with both data management systems. For instance, Reactome is a repository of biological pathways and reactions available as a BioMart database [68]. However, the Reactome Data Sharing Project, an extension of Reactome, aims to ensure that all data from this initiative will be fed into caBIG® [69].

Likewise, the caBIG® tool Cancer Models Database (caMOD) [70], which provides information about animal models for human cancers, has links to the Rat Genome Database [71], which is available as a BioMart resource [72]. These additional provisions by Reactome and the Rat Genome Database ensure increased interoperability with other resources, maximal dissemination of data and sustainability of the resource [73].

As the volume of data generated by the TCGA and the ICGC becomes accessible, via incorporation into the public repositories, GEO or ArrayExpress, certain statistical challenges, such as those attributed to the small number of samples or noise distribution across different platform-types, could be overcome to maximise the impact of the research generating the data.

It is important that researchers be made aware of potential pitfalls encountered when attempting to exploit data derived from different platforms and studies. For instance, data normalisation and integration becomes troublesome when conducting cross-platform meta-analyses. Indeed, there are no tools available for normalisation of data from different platforms. As such, data normalised separately are not directly comparable.

Furthermore, study design bias needs to be avoided. For example, if comparing cancer expression profiles to normal profiles, it is important to note the origin of the study. If all the cancer data are obtained from one study and all control data are from another, it is possible that any differences observed are attributable to study design rather than cancer.

The success of taking advantage of any data integration technology depends on a biological model assumption that allows for meaningful understanding of cancer development mechanisms.

Further problems in identification and integration of data are attributed to the use of differing terminologies between studies. In an attempt to tackle this problem, caBIG® has formulated standard controlled vocabularies to be adopted by the user-community. This ensures homogeneity in the terminology being used, thereby facilitating the data integration process.

From this review, it becomes apparent that for seamless information connectivity and optimal data sharing, the cancer research community needs to be able to liberate and unify the cancer data. Compelling researchers to design their initiatives to adhere to a standard infrastructure will meet with resistance and will probably fail in the short term. To this end, we propose that novel resources be designed to be interoperable with major international efforts in cancer research, such as the ICGC and the TCGA. That said, some level of compliance with widely applied infrastructures, such as the BioMart schema or compatibility with caBIG®, appears a necessary step to facilitate the integration of a novel sustainable resource. This in turn will allow the design and implementation of more sophisticated analysis portals.

In order to assist with this bioinformatics revolution, there is the requirement for a seismic shift in cancer research philosophy, with the criteria by which academic commendation is achieved changing from an institutionalised view to a community focus.

Key Points

- The heterogeneity and isolation of cancer data contributes to the segregation of vital research.
- Major national and international research efforts are being dedicated to the unification of these silos of information.
- There is a requirement for user-friendly interoperable integration systems that allow for the data to be accessed and analysed as a coherent whole.
- It is vital that novel resources are interoperable with existing infrastructures.

Acknowledgement

We thank Dr Arek Kasprzyk from the Ontario Institute for Cancer Research (OICR), Dr Alan Hogg and Dr Stuart Bell from the National Cancer Research Institute (NCRI) Informatics Initiative and Dr Carl Schaefer from the National Cancer Institute (NCI) for the critical reading of this review.

FUNDING

Cancer Research UK programme (C355/A6253).

References

1. von Eschenbach AC, Buetow K. Cancer Informatics Vision: caBIG. *Cancer Inform* 2007;**2**:22–4.
2. Lu Q, Hao P, Curcin V, *et al.* KDE Bioscience: platform for bioinformatics analysis workflows. *J Biomed Inform* 2006;**39**: 440–50.
3. Saltz J, Oster S, Hastings S, *et al.* caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. *Bioinformatics* 2006;**22**:1910–16.
4. Buetow KH. Cyberinfrastructure: empowering a "third way" in biomedical research. *Science* 2005;**308**:821–4.
5. Buetow KH. An infrastructure for interconnecting research institutions. *Drug Discov Today* 2009;**14**:605–10.
6. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 2000;**100**:57–70.
7. Hanash S. Integrated global profiling of cancer. *Nature Reviews* 2004;**4**:638–43.
8. The Cancer Genome Atlas. http://cancergenome.nih.gov (25 February 2010, date last accessed).
9. Hanauer DA, Rhodes DR, Sinha-Kumar C, *et al.* Bioinformatics approaches in the study of cancer. *Curr Mol Med* 2007;**7**:133–41.
10. Nevins JR, Potti A. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nat Rev Genet* 2007;**8**:601–9.
11. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008; **455**:1061–8.
12. Cancer Genome Workbench. http://cgwb.nci.nih.gov (25 February 2010, date last accessed).
13. UCSC cancer genome browser. https://genome-cancer.ucsc.edu (25 February 2010, date last accessed).
14. Memorial Sloan-Kettering Cancer Center (MSKCC). http://cbio.mskcc.org/cancergenomics-dataportal (25 February 2010, date last accessed).
15. Cancer Molecular Analysis portal (CMA). http://cma.nci.nih.gov (25 February 2010, date last accessed).
16. Pathway Interaction database. http://pid.nci.nih.gov/ (25 February 2010, date last accessed).
17. Schaefer CF, Anthony K, Krupa S, *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res* 2009;**37**:D674–9.
18. International Cancer Genome Consortium. www.icgc.org (25 February 2010 last accessed).
19. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009;**458**:719–24.
20. The Cancer Genome Project. www.sanger.ac.uk/genetics/CGP/ (25 February 2010, date last accessed).
21. Toner B. 'ICGC will use federated approach to manage impending crush of cancer genome data. *BioInform* May 2008. http://www.genomeweb.com/informatics/icgc-will-use-federated-approach-manage-impending-crush-cancer-genome-data (17 March 2010, date last accessed).
22. BioMart. http://www.biomart.org (25 February 2010, date last accessed).
23. Haider S, Ballester B, Smedley D, *et al.* BioMart Central Portal—unified access to biological data. *Nucleic Acids Res* 2009;**37**:W23–7.
24. Smedley D, Haider S, Ballester B, *et al.* BioMart—biological queries made easy. *BMC Genomics* 2009;**10**:22.
25. Spjuth O, Helmus T, Willighagen EL, *et al.* Bioclipse: an open source workbench for chemo- and bioinformatics. *BMC Bioinformatics* 2007;**8**:59.
26. Giardine B, Riemer C, Hardison RC, *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 2005;**15**:1451–5.
27. Cline MS, Smoot M, Cerami E, *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2007;**2**:2366–82.
28. Hull D, Wolstencroft K, Stevens R, *et al.* Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* 2006;**34**:W729–32.
29. Liu X, Wu J, Wang J, *et al.* WebLab: a data-centric, knowledge-sharing bioinformatic platform. *Nucleic Acids Res* 2009; **37**:W33–9.
30. Ruby API. http://github.com/dazoakley/biomart/ (25 February 2010, date last accessed).
31. Durinck S, Moreau Y, Kasprzyk A, *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 2005;**21**:3439–40.
32. Bioconductor. http://www.bioconductor.org/ (25 February 2010, date last accessed).

33. Dowell R, Jokerst R, Day A, *et al*. The Distributed Annotation System. *BMC Bioinformatics* 2001;**2**:7.

34. Ensembl. http://www.ensembl.org (25 February 2010, date last accessed).

35. The Pancreatic Expression database. http://www.pancreas expression.org (25 February 2010, date last accessed).

36. Chelala C, Hahn SA, Whiteman HJ, *et al*. Pancreatic Expression database: a generic model for the organization, integration and mining of complex cancer datasets. *BMC Genomics* 2007;**8**:439.

37. Chelala C, Lemoine NR, Hahn SA, *et al*. A web-based platform for mining pancreatic expression datasets. *Pancreatology* 2009;**9**:340–3.

38. Rossille D, Burgun A, Pangault-Lorho C, *et al*. Integrating clinical, gene expression, protein expression and preanalytical data for in silico cancer research. *Stud Health Technol Inform* 2008;**136**:455–60.

39. Cancer Biomedical Informatics Grid. https://cabig.nci.nih .gov (25 February 2010, date last accessed).

40. Cimino JJ, Hayamizu TF, Bodenreider O, *et al*. The caBIG terminology review process. *J Biomed Inform* 2009;**42**: 571–80.

41. Stein LD. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nat Rev Genet* 2008;**9**:678–88.

42. Madhavan S, Zenklusen JC, Kotliarov Y, *et al*. Rembrandt: helping personalized medicine become a reality through integrative translational research. *Mol Cancer Res* 2009;**7**: 157–67.

43. The Lymphoma Enterprise Architecture Data-system. http://umlmodelbrowser.nci.nih.gov/umlmodelbrowser/ (25 February 2010, date last accessed).

44. Huang TS, Sinha PJ, Graiser R, *et al*. Development of the Lymphoma Enterprise Architecture Database: a caBIG (tm) silver level compliant system. *Cancer Inform* 2009;**3**: 45–64.

45. Repository of Molecular Brain Neoplasia Data. https:// caintegrator.nci.nih.gov/rembrandt/ (25 February 2010, date last accessed).

46. ONIX. http://www.ncri-onix.org.uk (25 February 2010, date last accessed).

47. NCRI Informatics Initiative. http://cancerinformatics .org.uk/ (25 February 2010, date last accessed).

48. McCusker JP, Phillips JA, Beltran AG, *et al*. Semantic web data warehousing for caGrid. *BMC Bioinformatics* 2009;**10**:S2.

49. Beltran AG, Finkelstein A, Wilkinson JM, *et al*. Domain concept-based queries for cancer research data sources. *In: Proceeding of the Twenty-second IEEE International Symposium on Computer-Based Medical Systems*. CBMS 2009 Albuquerque, New Mexico, USA.

50. Ball CA, Brazma A, Causton H, *et al*. Submission of microarray data to public repositories. *PLoS Biol* 2004;**2**:E317.

51. Zhu Y, Davis S, Stephens R, *et al*. GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics* 2008;**24**:2798–800.

52. Barrett T, Troup DB, Wilhite SE, *et al*. NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res* 2007;**35**:D760–5.

53. Savage CJ, Vickers AJ. Empirical study of data sharing by authors publishing in PLoS Journals. *PLoS ONE* 2009;**4**: e7078.

54. Anderle P, Duval M, Draghici S, *et al*. Gene expression databases and data mining. *Biotechniques* 2003;**Suppl**: 36–44.

55. Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/ geo/ (25 February 2010, date last accessed).

56. Barrett T, Troup DB, Wilhite SE, *et al*. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* 2009;**37**:D885–90.

57. Barrett T, Edgar R. Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*. *Methods Mol Biol* 2006;**338**:175–90.

58. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol* 2006;**411**:352–69.

59. Sean D, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 2007;**23**:1846–7.

60. ArrayExpress. http://www.ebi.ac.uk/microarray-as/ae/ (25 February 2010, date last accessed).

61. Parkinson H, Kapushesky M, Kolesnikov N, *et al*. ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* 2009;**37**:D868–72.

62. Brazma A, Parkinson H, Sarkans U, *et al*. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2003;**31**:68–71.

63. Kauffmann A, Rayner TF, Parkinson H, *et al*. Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics* 2009.

64. Gene Expression Atlas. http://www.ebi.ac.uk/gxa/ (25 February 2010, date last accessed).

65. Oncomine. https://www.oncomine.org (25 February 2010, date last accessed).

66. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, *et al*. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* 2007;**9**:166–80.

67. MacDonald JW, Ghosh D. COPA—cancer outlier profile analysis. *Bioinformatics* 2006;**22**:2950–1.

68. Reactome BioMart Portal. http://www.reactome.org: 5555/biomart/martview/ (25 February 2010, date last accessed).

69. Reactome data sharing project. https://cabig.nci.nih.gov/ tools/Reactome (25 February 2010, date last accessed).

70. Cancer Models Database. http://cancermodels.nci.nih.gov (25 February 2010, date last accessed).

71. Rat Genome Database. http://rgd.mcw.edu/ (25 February 2010, date last accessed).

72. Rat Genome Database BioMart Portal. http://biomart.mcw .edu:9999/biomart/martview (25 February 2010, date last accessed).

73. de BonoB VI, D'Eustachio P, *et al*. Reactome: an integrated expert model of human molecular processes and access toolkit. *J Integr Bioinform* 2007;**4**(3):84–95.