Bioinformatics Integration for Cancer Research- Goal Question analysis

Leah Goldin
Software Engineering Department
Afeka-Tel Aviv College of Engineering
Tel-Aviv, Israel
leah@afeka.ac.il

Abstract **The PRM (Platform Reference Model) project proposes a model that enables cancer researchers to systematically access and use existing bioinformatics repositories.. Currently, there are many bioinformatics resources and standards both in the UK and internationally that are excellent research tools, but they have evolved separately and so present an incoherent and fragmented landscape.**

The Platform will provide access to an organised collection of informatics resources (applications and databases) for the acquisition, storage and use of data across the whole spectrum of cancer research.

This paper will focus on the requirements analysis process, aiming to understand the underlying concept of bioinformatics integration from a research goal-question viewpoint. Modeling a cancer research via research goal-question enables a planned collection of bioinformatics data along with a systematic approach of analyzing the reports answers to the research questions and goals.

***Index Terms*— Bioinformatics, Requirement analysis**

I. Introduction

As of today, cancer researchers use much existing bioinformatics resources in order to design, analyze and evaluate their experiments or studies. Traditional biomedical repositories of published papers are used by researchers to find results of successful experiments, relate their new goals for experiments to the previous ones, and use the published information in order to improve their experiment planning and design. More modern bioinformatics repositories, like EBI databases, contain genes expression variation, pathways analysed data, etc., that researchers use to analyze their experiments outcomes. Once performing registration to these repositories, every researcher is free to access them one by one. The reference platform will serve as an important vehicle to ease the researchers' access to the bioinformatics repositories in terms of time and efficiency. Moreover, we aim to improve the effectiveness of researchers access to these repositories, by providing integrative bioinformatics asset, which is retrieved from several bioinformatics resources based on the research questions (goals) provided by the researcher.

For example, if the researcher would have a research question concerning repositories {A, B, C} the PRM will be able to provide an integrated reference to all 3 repositories, enabling him to perform a more inclusive analysis, rather than accessing repositories {A, B} separately, and may be forgetting repository {C}.

A. Bioinformatics for Cancer Research

Currently, there are many bioinformatics resources and standards both in the UK and internationally. Many of these are excellent research tools, but they have evolved separately and so present an incoherent, fragmented landscape. At present there is scattered funding for these cancer projects, however, there is a need for integration and the problem is so large that a strategic direction and dedicated funding are required to integrate existing resources and build what is missing to create ac informatics platform for cancer research.

The goal of the NCRI [3] Informatics Initiative is to increase the impact of the UK cancer research and improve prevention and treatment of cancer by effective use of informatics to manage and exploit the vast amounts of diverse information currently generated. The scope of the Initiative is the full spectrum from basic to clinical cancer research and the bridge into health service delivery.

B. What is the Bioinformatics Platform?

The concept of the Platform is an organised set of informatics resources. Many of the projected components of the Platform are represented in the planning matrix produced by the NCRI Coordination Unit. The Platform will provide access to an organised collection of informatics resources (applications and databases) for the acquisition, storage and use of data across the whole spectrum of cancer research. Many of the resources developed will have applicability outside cancer. Many of the component parts of the Platform exist in isolation; the vision of the Platform is to integrate these at the level of data and/or software architecture. The process of developing the Platform will identify and fill current gaps in resources and will enable data sharing, and facilitate integration within and between different research domains.

C. *Related Work*

Several organisations are already tackling some of the important cancer informatics issues. These include the US National Cancer Institute Centre for Bioinformatics (NCICB) [1] and the European Bioinformatics Institute (EBI) [5]. The NCRI [3] has established strong working relationships with both.

Relevant UK national initiatives include OnCore UK [6], the NCRN [7] /UKCRN [8] electronic Remote Data Capture (eRDC) project, the e-Science Programme [9] and the health service programmes for IT in England, Wales, Scotland and Northern Ireland. Linking these projects nationally and internationally will produce a platform that is greater than the sum of the parts.

D. *Structure of the Paper*

In Section 2 an overview is given of the Platform Reference Model (PRM) scope and users along with the cancer-research domain model. In Section 3 a case study is described in order to demonstrate the goal-question model of cancer research. In Section 4 the overall requirements analysis flow is described. Section 5 presents discussion and future work.

II. THE PLATFORM REFERENCE MODEL (PRM)

A. *The PRM Scope*

The Platform modeled by the PRM project will enable data sharing, and facilitate integration within and between different cancer research areas.

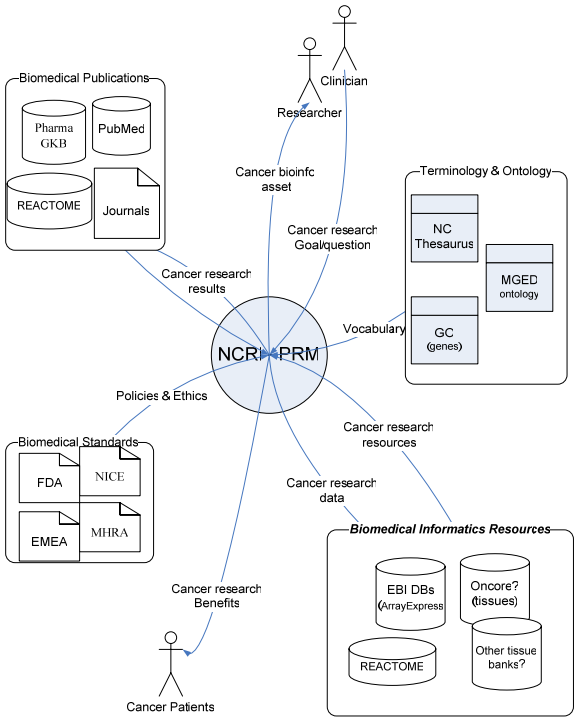

Figure 1: The PRM context diagram

The PRM context diagram in Figure 1 describes the interfaces between the PRM and its related external systems and subsystems; and how the PRM fits into the existing environment of cancer research informatics.

The external systems that interface with the PRM can be other systems such as informatics repositories, or humans such as researchers. The biomedical ***Terminologies & Ontologies*** include all the standard vocabularies needed to access the various informatics repositories, i.e., NCI [1] Thesaurus. ***Biomedical Informatics Resources*** include the existing informatics systems that contain raw data from different research areas, i.e., EBI databases such as, ArrayExpress. ***Researchers & Clinicians*** include researchers from different areas, i.e. biology, pathology, computer science, statisticians, etc., and clinicians, i.e. oncologist. ***Biomedical Publications*** include approved and validated research results published in medical journals, i.e., PubMed website, PharmaKGB database etc. ***Biomedical Standards*** include different policies and ethics procedures relevant to cancer research, such as FDA, NICE, EMEA, MHRA. ***Cancer Patients*** are people with cancer to which PRM will output the benefits of cancer research (directly or indirectly), i.e. transferring findings from cancer research into more effective treatment and care for cancer patients.

B. *The PRM Users*

The cancer researchers are of two types: scientific researchers and clinical researchers. The primary user of the PRM will be the scientific researcher. There are a number of different types of scientific researcher that will benefit form the Platform: Biologists, Chemists, Physicists, Statisticians Bioinformaticians, Computer Scientist, will use the Platform to support their experiments or studies. The PRM will enable the acquisition, storage, sharing and analysis of research and should improve the effectiveness and efficiency of scientific investigation.

Clinical Researcher, i.e., Radiologists, Pathologists, Oncologists (medical and surgical), might use the PRM to support a clinical trial, in order improve the trial effectiveness and efficiency, while making his/her results and analysis available via the PRM for future usage.

C. *The Cancer-Research Domain Model*

A domain model is a conceptual model representing the knowledge in a specific domain In the PRM we have identified two domain models: the cancer-biology domain model and the cancer-research domain model. The cancer-biology model is not in the PRM project scope, and is adopted from other international projects such as NCI [1] MGED [2].

The cancer-research domain model includes entities such as experiments, protocols, clustering services, etc. relevant to the cancer research itself, while the cancer-biology model includes entities such as genes, pathways, etc. relevant to biology in general and involved in cancer biology specifically.

The central element of the cancer-research model is the overall concept of *Investigation* which is defined as a study that examines some high level biomedical question, i.e., the role of diet in cancer, or whether a disease responds to a drug. An *experiment* then takes as input some biomedical

Resources, and is executed according to a *Protocol* and produces some *Results*.

Different kinds of experiments may require very different kinds of *resources*, including blood samples, mice, statistical data, [anonymous] patient information, etc. *IndividualInformation* can be found in specific registries of samples, usually stored in tissue banks.

Protocol represents the experiment designed and approved (if approval is needed) by ethical committee or other bodies in charge of the experiment authorization. *ProtocolApplication* describes how the protocol of an experiment has been actually implemented during the experiment execution. The actual protocol should include eventual variations to the approved protocol that, however, should not impact the adherence to the *Policies* which rule the experiment itself.

Finally, *ExperimentResults* model all the results produced by the experiment, including data published in curetted repositories and in scientific papers.

III. The Goal-Question Investigation Model

As we witnesses the ever-increasing amounts of cancer research data are collected and recorded in non-standardised ways and are not in a suitable form for sharing, re-use and integration. Thus, opportunities to gain new knowledge are lost, results are not translated for clinical use and experiments are repeated wastefully.

The NCRI platform initiative [3] vision is to increase the impact of UK cancer research and improve cancer prevention and treatment by creating interoperation and means for organising and managing research investigations.

The PRM aims to provide a platform model to support cancer research interoperation, i.e., support for planning and analyzing investigations through a form of workflow, support for resource discovery, support for recording and auditing investigations (for repeatability), access control and checks to avoid violation of ethical and other policies, data provenance and consistency checks.

The Goal-Question Report (GQR) method described in figure 2 is based on the traditional Goal-Question-Metric (GQM) method used in software measurements [4]. Planning measurement via GQM enables a planned collection of data along with a systematic approach of analyzing the metrics in order to get an effective managerial feedback.

The Goal Question Report (GQR) method rational is:

1. define your research goal,
2. state the research questions that are asked in order to check if the goal is met, and
3. identify the reports that will be used to answer these questions.

Planning cancer investigation / research in this way enables a planned collection of bioinformatics data along with a systematic approach of analyzing the reports answers to the research questions, in order to get an effective research result (wrt research goal).

In general the GQR investigation model can be used as a mechanism for both organizing and checking the biomedical informatics needed to support a cancer research. The GQR model is in the first place useful for designing a well defined research via goal/question modeling in order to achieve effective research analysis of outcomes. Once the research is managed via GQR, it can be used as guidance for checking the required bioinformatics resources.

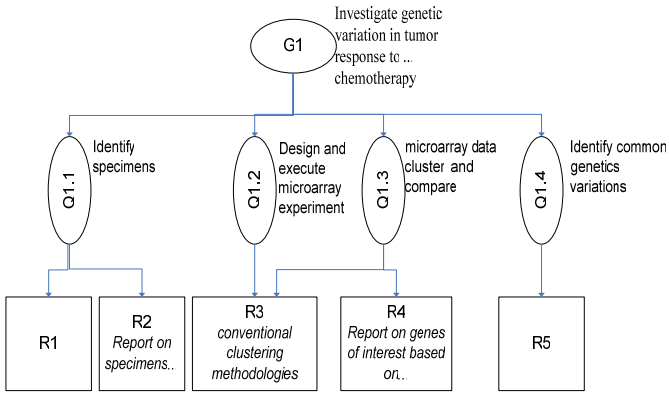

Figure 2: GQR method

We also present a cancer research case study analysis in order to demonstrate the bioinformatics assets required by the cancer researcher in order to carry out a successful research.

A. *Case-study - GQR Analysis*

The GQR analysis of the "Genetic Variation in Response to chemotherapy" case study in figure 3 includes the identification of research goal and questions, along with the required reports mapped to the existing Bioinformatics repositories.

The goal of this case study is (G1) to *investigate genetic variation in tumor response to treatment with a specific class of chemotherapy*. The research questions derived by the cancer researcher include (Q1) *identification of specimens of a specific tumour type, flash-frozen and prepared using a specific methodology, and for which there are associated medical records for treatment outcome,* and (Q3) *comparison of the clusters to currently-known metabolic pathways, some of which are known to be chemotherapy targets.*

The PRM reporting required to answer Q3 include datasets such as (R4) *the list of genes from the pathways of interest showing expression variation related to the chemotherapy treatment*, and data services such as (R3) *analyze the data using conventional clustering methodologies* (R3).

Figure 3: The GQR analysis of the case study

This case study analysis resulted 1 research goal, 4 research questions and 5 reports (see also Figure 2):

G1→ {Q1, Q2, Q4, Q5}

Q1→ {R1, R2} | Q2 →{R3} | Q3→ {R3, R4} | Q4 → {R5}

B. *Case-study – GQR Bioinformatics Analysis*

As part of the GQR analysis of the "Genetic Variation in Response to chemotherapy" case study, we have accompanied each dataset or report required to answer a research questions with reference to existing bioinformatics repositories, see *Bioinfo map* in figure 4. For example, EBI/Arrayexpress [5] required for microaaray analysis of R3, see table I.

TABLE I. GQR REFERENCE TO BIOINFORMATICS REPOSITORIES

Report	**Data asset (report/service)**	**Bioinformatics repositories**
R3	Analyse [and store] microarray data using conventional clustering methodologies	R3 ← Journals(JCO,Nature) ← clustering tools → EBI/ArrayExpress → Pub-Med →GEO → caArray
R4	Report on genes of interest based on expression in samples, and presence in metabolic pathways of interest, which are known to be chemotherapy targets	R4 ← REACTOME ← KEGG ← Journals ← Pub-Med

IV. THE PRM ANALYSIS FLOW

The PRM analysis process was done in parallel via the domain analysis and investigation analysis, see figure 4.

A. *Domain Analysis*

The domain analysis was the initial activity done mainly with the NCRI Unit project coordinators – known to be cancer research experts.. Defining the domain model was an iterative activity, using case studies intertwined with discussions with domain experts to reach a common vision. Since this domain context is continuously evolving, the model needs had to be sufficiently generic and extendible.

Eventually the domain analysis produced the Research domain model identifying key entities of cancer investigation as described in Section 2, while having better understanding of the relation to the Biology domain modeled by the existing international ontologies such as NCI [1], MGED, etc.

However, not much progress was achieved regarding the Research domain model regarding metadata needed to reference the biomedical informatics resources, see the *domain-bioinfo map* in figure 4.

B. *Investigation Analysis*

Once the Domain model was introduced, the GQR method was used for cancer investigation analysis, producing an investigation model of a cancer research that tremendously clarified the cancer research scope and the researcher needs for bioinformatics.

Once the researcher presents a research question (goal), by using the GQR we analyzed the expected bioinformatics report(s) required to answer that specific research question. These reports were then referenced to the bioinformatics resources, and presented to the researcher as an integrated *Bioinfo asset*, see figure 4.

Recall, the GQR is a very high level analysis of the research needs, involving an experienced researcher that knows which bioinformatics resources he will be using to answer her research questions. However, we plan to have the PRM referencing capability to the bioinformatics repositories done somehow automatically via the cancer Research and Biology domain models.

C. *Bioinformatics Resources Mapping*

Once the Investigation analysis specified the context and scope of the research needs, the Domain model was applied to map the GQR reports to the existing bioinformatics repositories. At this point, it was quite clear how to proceed with a deeper domain analysis in order to reveal the domain models relations, *Domain-bioinfo map* in figure 4, required to map the GQR reports to the bioinformatics resources, see table II.

D. *Al-in-all Analysis*

As described in Figure 4, the two parallel plans of analysis, i.e., Domain and Investigation models, actually complement each others. The Domain model provides a conceptual mechanism to map the cancer research domain generic entities to the bioinformatics resources, while the Investigation model via GQR provides a cancer research scope that "guides" the Domain analysis mapping.

TABLE II. DOMAIN MODEL VS. BIOINFORMATICS REPOSITORIES

GQR report	Research domain model	Biology domain model	Bioinfo
R3	***Protocol***	***SurfaceType*** (polylysine, aminosilane) ***TechnologyType*** (spotted_antibody_features , spotted_ds_DNA_features)	GEO caArray ArrayExpress Pub-Med
	ExpResult	*DataTransformationProtocolType* **(**loess_group_normalization,loess_global_normalizo) *(MGED_Ontology)*	ArrayExpress caArray
R4	***Agent***	***Gene,Gene_Product, Pathway, Chemicals_and_Drugs*** ***(NCI_Thesaurus)***	PharmGKB Reactome Pub-Med

The missing link currently is the translation of the Investigation reports to the Research domain entities. This translation will use the required reports of the GQR and transform them into the Bioinfo language by which the Research domain model (based on the Biology model) will then be used to reference the Bioinformatics resources.

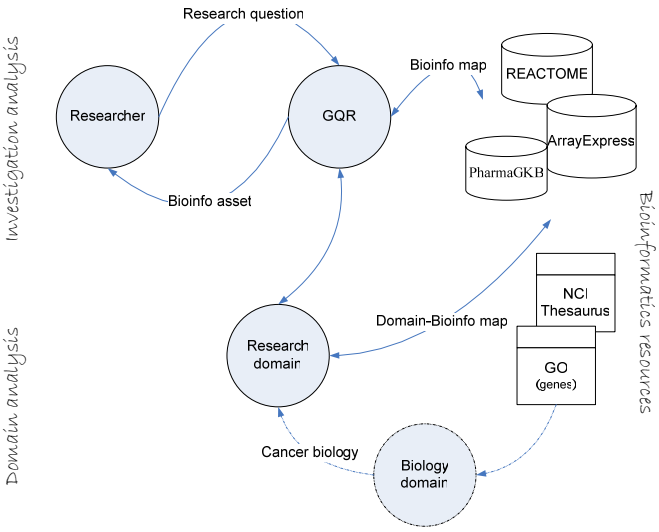

Figure 4: The PRM analysis flow

V. DISCUSSION AND FUTURE WORK

The PRM is a typical e-science project aiming to integrate existing bioinformatics resources. The aspects characterizing an e-science project include heterogeneous information, multidisciplinary teams, intrinsically distributed systems and cross-organization communication, changing of context, etc. In the PRM project we have already faced many of these aspects, and are using different modeling techniques in order to capture as many as possible viewpoints.

The GQR method was used for cancer research analysis, producing an investigation model of a cancer research that clarified the cancer research scope and the researcher needs for bioinformatics. Planning cancer research via GQR enables a planned collection of bioinformatics data along with a systematic approach of analyzing the reports answers to the research questions and goals.

One of the initial steps in the project was to create the PRM stakeholders model which is not described in here because of space limitations. This list of PRM stakeholders is used as a checklist to identify the origins of requirements of the PRM. Unfortunately it was very difficult to meet representatives of these stakeholders. Even those we met eventually did not have much experience with using bioinformatics resources, and only few researchers could envision the future PRM capabilities and were mainly concentrated in current technical difficulties with computers.

Using different case studies relevant to different cancer research areas became a crucial resource for analyzing the PRM requirements and validating the PRM models.

The main challenge is to analyze the concept of bioinformatics integration via content-wise point of view, i.e., what is the content of the “bioinformatics integration” needed by the cancer researcher?

(1) Acknowledgment

This work was done at University College London (UCL) UK, under the supervision of Prof. Anthony Finkelstein the Head of the Computer science Department, supported by the NCRI Informatics Initiative.

REFERENCES

[1] NCICB: US National Cancer Institute Centre for Bioinformatics. www.ncicb.nci.nih.gov

[2] MGED: Microarray Gene Expression Data Society http://www.mged.sf.net/ontologies

[3] NCRI, National Cancer research Institute www.cancerinformatics.org.uk

[4] .Solingen, E.Berghout: *The Goal/Question/Metric Method*, McGraw-Hill Publishing Company, 1999

[5] EBI, European Bioinformatics Institute. http://www.ebi.ac.uk/

[6] OnCore, previously NCRI National Cancer Tissue Resource

[7] NCRN, National Cancer Research Network www.ncrn.org.uk

[8] UKCRN, UK Clinical Research Network www.ukcrn.org.uk

[9] eRDC, NeSC- National e-Science Centre. http://www.nesc.ac.uk/