European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

Deliverable D4.8 Feature Roadmap 3

Due date of deliverable: Month 30

Actual submission data: Month 31

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including Commission Services)	
RE	Restricted to a group specified by the consortium (including Commission Services)	
CO	Confidential, only for members of the consortium (including Commission Services)	

DELIVERABLE INFORMATION

Project	
Project acronym:	eTRIKS
Project full title:	European Translational Information and Knowledge Management Services
Grant agreement no.:	115446
Document	
Deliverable number:	D4.8
Deliverable title:	Feature Roadmap 3
Deliverable version:	1
Due date of deliverable:	Month 30
Actual submission date:	Month 31
Leader:	
Editors:	
Authors:	Mansoor Saqi, Venkata Satagopam, Manfred Hendlich, Wei Gu
Reviewers:	Chrish Marshall, Milan Ganguly
Participating beneficiaries:	CNRS, BioSci
Work Package no.:	WP4
Work Package title:	Analytics Research & Content Curation
Work Package leader:	Manfred Hendlich and Reinhard Schneider
Work Package participants:	
Estimated person-months for deliverable:	1
Nature:	PU
Version:	1
Draft/Final:	Final
No of pages (including cover):	8
Keywords:	data analytics, visualisation

Overview

This document builds on the roadmap D4.6, and describes the status of eTRIKS with respect to the roadmap, identifying areas where further development is needed. These include: better visualisation, robust workflows for some common analytic tasks, development of detailed use cases that highlight the features of the platform and further development of a bio-centric knowledge base.

Scope of this document

Deliverable D4.6 (07/14) described the eTRIKS analytics and visualisation roadmap. The current document (D4.8) describes how developments to the eTRIKS platform are aligned with the roadmap and identifies future directions. The WP4 team will add to, and update this document on a regular basis.

Contents

Introduction

The overall vision outlined in D4.6 was summarised in Figure 1, which is displayed again at the end of this document. The main themes identified in the roadmap were

(a) Core tranSMART developments

(b) Development of a few robust basic analysis tools and a mechanism for easy connection to external tools, case studies and user stories to showcase the functionality in the platform and

(c) Development of a knowledge base to put results from translational research experiments into biological context and facilitate hypothesis generation.

We note that many IMI projects are still in the stage of data collection, and early requirements are centred on curation and organisation of the data rather than on analytic features. As the on-going projects progress, we would envisage this situation to change. An exception is the IMI U-BIOPRED (Unbiased BIOmarkers in PREDiction of respiratory disease outcomes) project where data collection and analysis is in a more advanced stage, and requirements for this project have shaped some of the ideas presented here on data analytics for eTRIKS.

Core Functionality

Core functionality refers to functionality with the tranSMART platform that would facilitate a better environment for data analytics and includes API development better visualisation components with the platform, data provenance and security, and the ability of accommodate multiple high dimensional data types. These issues require close interaction between WP2 and WP4.

R client

This has been developed and is described in document D2.6 (in 'Met Requirements' Section 3). The R client allows secure connection to tranSMART and enables data to be received from tranSMART. The R client is being used and tested in different projects and a list of additional improvements for performance usability and functionality for the next phase of development, based on these experiences will be compiled. This includes the ability to select subsets that have already been saved using the tranSMART web application.

Better visualisation in tranSMART

Deliverable D4.6 identified the importance of better visualisation in tranSMART for new data types eg genetic variant information. The platform currently has limited visualisation capabilities compared to other tools (such as cBioportal or NextBio). The cBioportal OncoPrint component (which allows easy visualisation of mutations and other genomic changes across a set of samples) was identified as an example of a particularly useful and user-friendly tool for visual mining of complex data sets (refer to www.cbioportal.org). Discussions have taken place on establishing a collaboration involving cBioportal (MSKCC) and the Hyve/TraIT consortium. It is planned to engage further with academic and commercial groups who are developing visualisation components for tranSMART. Other approaches could include development of APIs to specialist visualisation tools. Part of the activities of a new eTRIKS project member at LCSB (1st quarter 2015) will be involvement in these collaborations. Developments to the platform will be done together with WP2.

Towards better Data Provenance:

Business rules for capturing and managing data provenance information in eTRIKS were defined in deliverable D4.3 'Data Provenance Guidelines'. Technical requirements for managing and displaying data provenance information within the "eTRIKS Harmonization" system will be provided to the next development phase of eTRIKS version 3.0. It is planned that the technical requirements will be collected by the work group of the "eTRIKS Harmonization" system that involves UL and ICL colleagues as discussed in the Barcelona annual meeting. Close interactions will be developed with the eTRIKS harmonisation system (WP2 and WP3, D3.5) regarding implementation of data provenance. Additionally the provenance model described by the W3C will be explored.

Analytic Functionality

Workflows

Deliverable D4.6 identified the importance of a core set of analytic functionality to eTRIKS and development of core analysis workflows for some common types of analyses needed by the translational medicine community. Some prototyping of using tranSMART together with Galaxy as a system for tools integration has been carried out. Following cohort selection, an export tab in tranSMART allows export of data to Galaxy. Galaxy workflows for marker

selection from RNAseq, RNA array, methylation and miRNA data have been developed. An advantage of Galaxy workflows is that the underlying complexity of the computational process can be hidden from the users when the system is used in basic mode. More experienced users can however enable advanced mode if they want to change settings and parameters of the workflow. Better integration of differentially expressed genes to pathways and the inclusion of over-representation and enrichment analyses remain to be implemented.

In addition to marker selection, experimental workflows for identification of disease subtypes using, as input, multiple 'omics datasets were explored. Two approaches were implemented, a graph based approach (Similarity Network Fusion[1]) and an ordination-based approach (Co-inertia Analysis[2]). Workflows for these two methods were implemented using Galaxy. Further work remains to be done on assessing the robustness of the workflows and on envisioning the results.

Further identification of common workflows based on specific use cases need to be developed and tested. This includes marker selection on small genomic variants, mRNA microarray, methylation microarray, RNA sequencing data (mRNA, miRNA)) These will also include sample based workflows that Oncotrack users need to track xenografts that requires multi-layer mapping between patient, patient sample, model sample(e.g. xenograft samples). Other requirements include functionality for cross study analysis. It is also recognised that in order to support more data types some performance issues in tranSMART will need to be addressed. A meeting is planned including participants from WP4 WP2, WP6, members of CTTM/trait and the Hyve to identify requirements relating to 'omics data features is scheduled for May 6th, 2015.

Towards more efficient analyses of high dimensional data.

The retrieval of high dimensional data such as gene expression data from tranSMART and construction of a data matrix (needed for analyses such as marker selection) is time intensive and can be a bottleneck in the total time taken to perform some analyses.

In an effort to explore alternative approaches to improve performance, a prototype system based on a noSQL database has been developed and preliminary results show significant improvements. The prototype system has used a document database for storing high dimensional data. Cohort selection is performed within tranSMART which is then connected with the database and the associated data collected as a data matrix. This can then be used in subsequent analyses. Within the document database new annotation types for each gene can easily be added (for example the pathways in which the gene participates).

Further work is needed to make the coupling of noSQL database with tranSMART seamless and more robust, and to ensure that the process is transparent from the perspective of the user.

Detailed Use Cases

Detailed use cases that show how the platform is used in a non trivial data analysis task need to be developed. These should include scenarios that focus on

(a) Exploration of clinical attributes

(b) Biomarker identification with integration to detailed pathway annotation

(c) Unsupervised methods for exploratory analysis of high dimensional data from multiple platforms

(d) Examples illustrating the use of the Biocentric knowledge base, for example to gain further insight into a set of differentially expressed genes emerging from a biomarker study.

Development of a Biocentric Knowledge Base

The use of a graph database (neo4j) for representation of background biological information such as protein interactions, drug-target relationships, disease-gene relationships and protein-pathway relationships has been explored. This database allows known disease genes or results from experimental studies, to be mapped to a network of background knowledge and facilitates identification of biological context.

The Minerva platform at LCSB allows results from experimental studies such as gene expression to be mapped to disease pathways (represented in SBML) and visualised. An exploratory study that links the Minerva technology to tranSMART has been carried out.

Exploration of methods of linking tranSMART to disease maps and the contextual knowledge base will be further developed. Exploratory discussions with the NDex group have been initiated and links with other on-going community curation tools and infrastructures need to be explored. The suitability of openBEL as a framework for knowledge representation should be assessed.

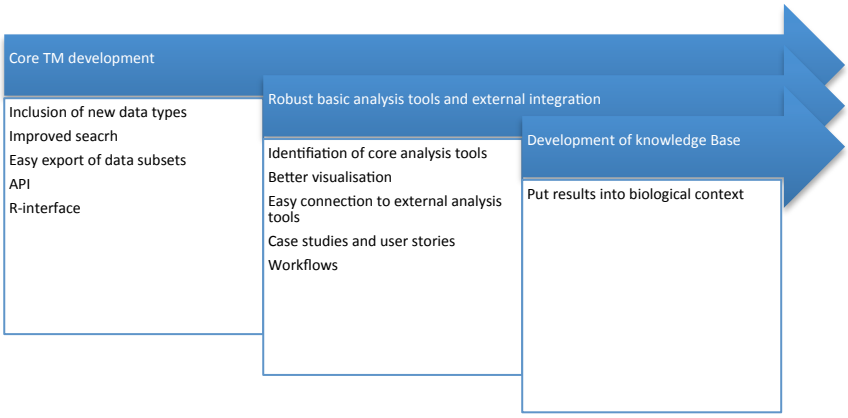

Figure 1 (taken from D4.6): Steps in the development of the analytics functionality in eTRIKS

References

[1] Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods *11*, 333–337

[2] Meng, C., Kuster, B., Culhane, A.C., and Gholami, A.M. (2014). A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics *15*, 162