



LarKC

The Large Knowledge Collider

a platform for large scale integrated reasoning and Web-search

FP7 – 215535

D7b.3.3a Version 3 iteration report

Coordinator: Mark A. Greenwood

With contributions from: Mark A. Greenwood, Niraj Aswani, Angus Roberts, Raluca Brehar, Mattias Johansson, James McKay, Jon Wakefield, Valentin Tablan, Ian Roberts, Hamish Cunningham, Paul Brennan

Quality Assessor: Bosse Andersson

Quality Controller: Mark A. Greenwood

Document Identifier:	LarKC/2008/D7b.3.3a/V1.0
Class Deliverable:	LarKC EU-IST-2008-215535
Version:	version 1.0
Date:	September 23, 2011
State:	final
Distribution:	public



EXECUTIVE SUMMARY

The Large Knowledge Collider (LarKC) project is building a platform for scaleable reasoning over terabytes of scientific data, using massive distributed incomplete reasoning. One of the use cases is carcinogenesis research, as described in LarKC Deliverable D7b1.1a *Requirements summary*. One scenario of this use case uses literature knowledge mining to assist with predicting gene-disease associations in Genome Wide Association Studies (GWAS).

During the life of the LarKC project, we have built prototype software that uses LarKC to assist with the GWAS scenario. This is Version 3 of the use case software, and is described in LarKC Deliverable D7b.3.3b *Version 3 prototype*.

This document gives a report of the use of the prototype software. First, a quantitative proof-of-principle evaluation of the software is given, showing that it makes a significant difference in the ranking of gene-disease associations. Secondly a qualitative evaluation of the interface by end-users is presented which is followed by a brief analysis of the keywords used by the software for text mining. Lastly a brief analysis of the system performance generated via the instrumented platform is reported.



DOCUMENT INFORMATION

IST Project Number	FP7 – 215535	Acronym	LarKC
Full Title	The Large Knowledge Collider: a platform for large scale integrated reasoning and Web-search		
Project URL	http://www.larkc.eu/		
Document URL			
EU Project Officer	Stefano Bertolo		

Deliverable	Number	7b.3.3a	Title	Version 3 iteration report
Work Package	Number	7b	Title	Carcinogenesis reference production







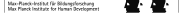




Date of Delivery	Contractual	M42	Actual	30-Sep-11
Status	version 1.0		final <input checked="" type="checkbox"/>	
Nature	prototype <input type="checkbox"/> report <input checked="" type="checkbox"/> dissemination <input type="checkbox"/>			
Dissemination Level	public <input checked="" type="checkbox"/> consortium <input type="checkbox"/>			

Authors (Partner)	Mattias Johansson, Paul Brennan, James McKay (all International Agency for Research on Cancer); Jon Wakefield (University of Washington); Niraj Aswani, Mark A. Greenwood, Ian Roberts, Valentin Tablan, Hamish Cunningham, Angus Roberts (all University of Sheffield); Raluca Brehar (Technical University of Cluj-Napoca)			
Resp. Author	Mark A. Greenwood		E-mail	m.greenwood@dcs.shef.ac.uk
	Partner	University of Sheffield	Phone	+44 (114) 222 1800

Abstract (for dissemination)	<p>Given advances in human genome sequencing, genetic testing, and the availability of samples from large population studies, it is now possible to carry out new types of study on the association between genes and diseases - Genome-Wide Associations Studies. In these, samples are tested from thousands of subjects with the disease in question, and thousands of disease-free controls. Each sample is tested with many hundreds of thousands of gene markers. If a marker is found more frequently in disease samples as opposed to control samples, then perhaps genes close to that marker are associated with the disease. Analysis of raw experimental data uses common statistical models to find the relevance of each marker, and rank them in order of relevance to the disease. Genes close to the top few markers are then studied in more depth. This last bit is expensive, and improving rankings could improve both the efficiency and the economics of the technique. Analysis could be improved if we incorporated knowledge we already have about genes - prior knowledge. Such knowledge is available e.g. in the vast numbers of research databases and research publications that now exist in the Life Sciences. This report evaluates a third version of the prototype software developed to assist with this. The prototype software was produced in the context of the Large Knowledge Collider (LarKC) project.</p>
Keywords	Evaluation, Genome Wide Association Study, GWAS, carcinogenesis, Bayesian statistics, Bayesian False Discovery Probability, Single Nucleotide Polymorphism, SNP, GeneRIF



PROJECT CONSORTIUM INFORMATION

Participant's name	Partner	Contact
Semantic Technology Institute Innsbruck, Universitaet Innsbruck	 	Prof. Dr. Dieter Fensel Semantic Technology Institute (STI), Universitaet Innsbruck, Innsbruck, Austria Email: dieter.fensel@sti-innsbruck.at
AstraZeneca AB		Bosse Andersson AstraZeneca Lund, Sweden Email: bo.h.andersson@astrazeneca.com
CEFRIEL - SOCIETA CONSORTILE A RESPONSABILITA LIMITATA		Emanuele Della Valle CEFRIEL - SOCIETA CONSORTILE A RE- SPONSABILITA LIMITATA Milano, Italy Email: emanuele.dellavalle@cefriel.it
CYCROP, RAZISKOVANJE IN EKSPERI- MENTALNI RAZVOJ D.O.O.		Michael Witbrock CYCROP, RAZISKOVANJE IN EKSPERIMEN- TALNI RAZVOJ D.O.O., Ljubljana, Slovenia Email: witbrock@cyc.com
Höchstleistungsrechenzentrum, Universitaet Stuttgart		Georgina Gallizo Höchstleistungsrechenzentrum, Universitaet Stuttgart Stuttgart, Germany Email : gallizo@hlrs.de
MAX-PLANCK GESELLSCHAFT ZUR FOERDERUNG DER WISSENSCHAFTEN E.V.		Dr. Lael Schooler, Max-Planck-Institut für Bildungsforschung Berlin, Germany Email: schooler@mpib-berlin.mpg.de
Ontotext AD		Atanas Kiryakov, Ontotext Lab, Sofia, Bulgaria Email: naso@ontotext.com
SALTLUX INC.		Kono Kim SALTLUX INC Seoul, Korea Email: kono@saltlux.com
SIEMENS AKTIENGESELLSCHAFT		Dr. Volker Tresp SIEMENS AKTIENGESELLSCHAFT Muenchen, Germany Email: volker.tresp@siemens.com
THE UNIVERSITY OF SHEFFIELD		Prof. Dr. Hamish Cunningham, THE UNIVERSITY OF SHEFFIELD Sheffield, UK Email: h.cunningham@dcs.shef.ac.uk
VRIJE UNIVERSITEIT AMSTERDAM		Prof. Dr. Frank van Harmelen, VRIJE UNIVERSITEIT AMSTERDAM Amsterdam, Netherlands Email: Frank.van.Harmelen@cs.vu.nl
THE INTERNATIONAL WIC INSTI- TUTE, BEIJING UNIVERSITY OF TECHNOLOGY		Prof. Dr. Ning Zhong, THE INTERNATIONAL WIC INSTITUTE Mabeshi, Japan Email: zhong@maebashi-it.ac.jp
INTERNATIONAL AGENCY FOR RE- SEARCH ON CANCER		Dr. Paul Brennan, INTERNATIONAL AGENCY FOR RE- SEARCH ON CANCER Lyon, France Email: brennan@iarc.fr
INFORMATION RETRIEVAL FACILITY		Dr. John Tait, Dr. Paul Brennan, INFORMATION RETRIEVAL FACILITY Vienna, Austria Email: john.tait@ir-facility.org





TECHNICAL UNIVERSITY OF CLUJ-NAPOCA http://www.utcluj.ro/	 The logo of the Technical University of Cluj-Napoca, consisting of a stylized 'T' and 'U' in red and grey.	Prof. Dr. Eng. Sergiu Nedevschi TECHNICAL UNIVERSITY OF CLUJ-NAPOCA Cluj-Napoca, Romania E-mail: sergiu.nedevschi@cs.utcluj.ro
SOFTGRESS S.R.L. http://www.softgress.com/	 The logo for Softgress, featuring the word 'Softgress' in white text on a blue rectangular background.	Dr. Ioan Toma SOFTGRESS S.R.L. Cluj-Napoca, Romania E-mail: ioan.toma@softgress.com



TABLE OF CONTENTS

LIST OF FIGURES	7
1 INTRODUCTION	9
1.1 Summary of the use case	9
1.2 Carcinogenesis use case iterations	9
1.3 Outline of the report	10
2 QUANTITATIVE PROOF-OF-PRINCIPLE EVALUATION	11
2.1 Introduction	11
2.2 Methods and rationale	11
2.3 Results	11
2.3.1 Subjectively assigned keywords	12
2.3.2 Priors assigned through random indexing	12
2.3.3 Priors assigned through TFIDF	12
2.3.4 Priors assigned through UMLS term expansion	13
2.3.5 Priors assigned through TFIDF and UMLS term expansion . . .	13
2.4 Summary	14
3 QUALITATIVE KEYWORD EVALUATION	17
4 USER EVALUATION	18
4.1 Introduction	18
4.2 Web service	18
4.3 Advantages	18
4.4 Limitations and areas of improvement	19
5 INSTRUMENTATION APPLIED ON WP7B WORKFLOWS	20
5.1 Instrumentation Results	20
6 CONCLUSION	24
REFERENCES	25



LIST OF FIGURES

5.1	CPU Usage and and Time	21
5.2	Amount of Free, Allocated and Memory Used	22
5.3	Query Level Instrumentation Results	23



LIST OF ABBREVIATIONS

B FDP	Bayesian False Discovery Probability
GWAS	Genome Wide Association Study
IARC	International Agency for Research on Cancer
LarKC	The Large knowledge Collider project
RI	Random Indexing
SNP	Single Nucleotide Polymorphism
T2D	Type 2 Diabetes
TFIDF	Term Frequency Inverse Document Frequency
UMLS	Unified Medical Language System
WHO	World Health Organisation



1. Introduction

This deliverable reports on the development of software to support the LarKC WP7b Carcinogenesis use case. The software is developed iteratively over the life of the LarKC project. This report covers the third and final iteration.

The first iteration of the software was delivered as *LarKC deliverable D7b.3.1b, Version 1 prototype* [1], and its development and evaluation reported in *LarKC deliverable D7b.3.1a, Version 1 iteration report* [8].

The second iteration of the software was delivered as *LarKC deliverable D7b.3.2b, Version 2 prototype* [2], and it was evaluated in *LarKC deliverable D7b.3.2a, Version 2 iteration report* [5].

The prototype software provides support for data analysis in Genome wide association studies (GWAS). When describing evaluations, the material reported assumes that the reader is familiar with GWAS. Simplified descriptions can be found in other deliverables. The use case is described in detail in *D7b.1.1a Requirements summary and data repository* [7], and a short, illustrated, description is given in *LarKC deliverable D7b.3.3b, Version 3 prototype* [4]. The following introduction to the use case is taken from the latter.

1.1 Summary of the use case

GWAS use bioprobes (SNPs - gene markers) to look for higher levels of association between genes in a diseased subjects as opposed to controls. The large numbers of markers mean that huge numbers of samples are needed to achieve sufficient statistical power.

Analysis of raw experimental data uses common statistical models to find the relevance of each marker, and rank them in order of relevance to the disease. Genes close to the top few markers are then studied in more depth. This last bit is expensive, and improving rankings could improve both the efficiency and the economics of the technique. Analysis could be improved if we incorporated knowledge we might already have about genes - prior knowledge. Such knowledge is available e.g. in the vast numbers of research databases and research publications that now exist in the Life Sciences, sometimes known as the data-ome and bibli-ome.

LarKC WP7b aims to apply LarKC technology to this problem, scaling knowledge discovery across the large amounts of biomedical knowledge now encoded in the data- and bibli-ome, and applying it to the millions of data points in a typical GWAS. We have prototyped a technique with the WHO's cancer research unit, IARC, to combine prior knowledge about a gene with experimental data, thus improving statistical power. The prototype uses early versions of LarKC plugins, and uses the LarKC data layer.

1.2 Carcinogenesis use case iterations

The general approach for reporting each use case prototype iteration is described in *D7b.1.1b Iteration evaluation methodology and report template* [6], which describes a system evaluation based on usability. By presenting quantitative evaluation results, we have departed from the usability evaluation plan. This departure is justified because:



- The usability evaluation gives a qualitative, user centric view. In this iteration, the software produces quantitative data. This can therefore be evaluated quantitatively - as is done in this report.
- The usability evaluation assumes a large amount of interaction between the user and an interface. In practice, there is very little end user interaction with the software. The interface is minimal. The software delivers data sets to the end-user, which they then manipulate in their existing tools.
- Usability has previously been shown by reserach papers presenting results from the previous iteration to end user conferences and workshops, given in [8].

1.3 Outline of the report

Three evaluations are presented. First, **Chapter 2** presents a quantitative examination of the software, by evaluating how it performs over SNPs already known to be associated with diseases. The second evaluation in **Chapter 3** gives a domain expert evaluation of example keywords generated by the system, and as used in knowledge mining. Third, **Chapter 4** gives a qualitative report of the system by end users. Finally, Chapter 6 draws some conclusions.



2. Quantitative proof-of-principle evaluation

2.1 Introduction

This section includes a quantitative evaluation from a proof-of-principle experiment aiming to evaluate if the priors produced by the GWA service improves ranking in genome-wide association studies.

2.2 Methods and rationale

In order to evaluate if the Bayesian false discovery probability (BFDP) method incorporating priors from the medical literature improves the ranking of SNPs in GWAS, we have conducted an experiment using results published on frequently studied endpoints, including prostate cancer, breast cancer, and type 2 diabetes (T2D). We have identified multiple SNPs that have been robustly associated with each endpoint, and assigned priors for each SNP using five methods:

1. subjectively assigned keywords;
2. random indexing;
3. TFIDF;
4. UMLS expansion;
5. TFIDF with UMLS expansion.

The prior distribution for the SNPs were subsequently compared to others included on commonly used genome-wide chips. The rationale for this approach is that if we can show that the priors for 'truly' associated SNPs of each respective endpoint are systematically higher than random SNP, it would directly imply that truly associated SNPs would be ranked higher in GWAS. This would provide a proof of principle experiment showing that the BFDP method works.

2.3 Results

We identified truly associated SNPs for prostate cancer ($n = 39$), breast cancer ($n = 19$), and Type 2 diabetes ($n = 35$) for in total 93 SNPs. Various methods of keyword selection were used to assign priors for all SNPs included on the commonly used Illumina 610 Quad genome-wide array. The distribution of the priors for all SNPs was compared to the distribution of the known truly associated SNPs to give a Chi-square statistic, and a p value. If a p value is significant for the method, this implies that the method leads to significantly higher ranking of truly associated SNPs. The following sections give detailed results for each method. A summary discussion and a summary of all p values are then given in Section 2.4 and Table 2.6.



2.3.1 Subjectively assigned keywords

In the subjectively assigned keywords experiment, the system used keywords assigned by domain experts.

The distributions of assigned priors are shown in Table 2.1. The priors were consistently higher for the truly associated SNPs than for random SNPs. The proportion of true SNPs in the bottom prior category for breast cancer was 37%, compared to 51% for random SNPs ($p = 0.004$), and similar differences were observed for Prostate cancer ($p = 4 \times 10^{-5}$) and T2D ($p = 2 \times 10^{-6}$).

Breast cancer			Prostate cancer			Type 2 Diabetes		
Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c	Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c	Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c
0.00004	51%	37%	0.00004	51%	31%	0.00002	78%	51%
0.00008	8%	0%	0.00007	7%	0%	0.00010	12%	9%
0.00009	21%	26%	0.00008	31%	54%	0.00023	6%	26%
0.00037	4%	16%	0.00066	1%	0%	0.00104	0%	0%
0.00113	17%	21%	0.00180	10%	15%	0.00288	4%	14%
p value ^d	Average prior for random SNPs	Average prior for true SNPs	p value ^d	Average prior for random SNPs	Average prior for true SNPs	p value ^d	Average prior for random SNPs	Average prior for true SNPs
0.004	0.0002	0.0003	4×10^{-5}	0.0002	0.0003	2×10^{-6}	0.0002	0.0005

Table 2.1: Distributions of priors based on subjectively assigned keywords

(a) The exact prior of each category was assigned based on (i) the observed distribution of random SNPs across the categories, and an (ii) assumed distribution and (iii) total number of truly associated SNPs. (b) The distribution was based on 598,805 SNPs included on the Illumina 610 quad genome-wide array. (c) True SNPs included SNPs previously robustly associated with breast cancer ($n = 19$), prostate cancer ($n = 39$), and Type 2 diabetes ($n = 35$). (d) The p -value is based on Chi-square statistics comparing the expected and observed distributions of 'truly' associated SNPs

2.3.2 Priors assigned through random indexing

In these experiments, the LarKC random indexing plugin was used to select keywords, when seeded with the name of the disease only.

The distributions of assigned priors are shown in Table 2.2. The differences in prior distributions between random and true SNPs were similar to those generated through subjective keywords, but a larger proportion of random SNPs were assigned to the highest prior group than in the experiment using subjective priors. The strongest difference in prior assignments between random and true SNPs were observed for T2D ($p = 4 \times 10^{-8}$), with 34% of true SNPs being assigned to the top prior category compared to 5% of random SNPs. Priors generated for breast and prostate cancer also showed notable differences in prior distributions ($p = 0.04$ and 0.001 , respectively), although to a lesser extent than for T2D.

2.3.3 Priors assigned through TFIDF

In these experiments, the standard information retrieval metric of Term Frequency Inverse Document Frequency was used to rank terms commonly found in the same



Breast cancer			Prostate cancer			Type 2 Diabetes		
Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c	Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c	Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c
0.00004	49%	32%	0.00004	36%	15%	0.00002	77%	49%
0.00008	3%	0%	0.00006	17%	0%	0.00011	12%	11%
0.00009	14%	21%	0.00009	32%	23%	0.00022	6%	6%
0.00025	14%	16%	0.00051	0%	0%	0.00091	0%	0%
0.00107	20%	32%	0.00128	16%	38%	0.00240	5%	34%
p value ^d	Average prior for random SNPs	Average prior for true SNPs	p value ^d	Average prior for random SNPs	Average prior for true SNPs	p value ^d	Average prior for random SNPs	Average prior for true SNPs
0.04	0.0003	0.0004	0.001	0.0003	0.001	4×10^{-08}	0.0002	0.001

Table 2.2: **Distributions of priors based on keywords generated through random indexing** (a) The exact prior of each category was assigned based on (i) the observed distribution of random SNPs across the categories, and an (ii) assumed distribution and (iii) total number of truly associated SNPs. (b) The distribution was based on 598,805 SNPs included on the Illumina 610 quad genome-wide array. (c) True SNPs included SNPs previously robustly associated with breast cancer ($n = 19$), prostate cancer ($n = 39$), and Type 2 diabetes ($n = 35$). (d) The p -value is based on Chi-square statistics comparing the expected and observed distributions of 'truly' associated SNPs

documents as the disease name. The highest ranking terms were used to generate priors. The technique is explained further in *LarKC deliverable D7b.3.2b, Version 2 prototype* [2].

Table 2.3 shows that priors assigned through TFIDF are significantly higher for true SNPs compared to random SNPs. For breast cancer and T2D, TFIDF places 2 and 3 times more true SNPs in a higher prior group, respectively. For prostate cancer, the technique places 5% of SNPs in the highest group, compared to 0% random SNPs.

2.3.4 Priors assigned through UMLS term expansion

In these experiments, the disease name was expanded to a set of terms related to the disease name in the UMLS Metathesaurus, as explained in *LarKC deliverable D7b.3.2b, Version 2 prototype* [2].

Table 2.4 shows that with breast cancer, UMLS expansion fails to make a significant difference in the priors of true SNPs. There is, however, a significant difference for both prostate cancer ($p = 3 \times 10^{-7}$) and T2D ($p = 0.0004$). For prostate cancer, the technique again places 5% of SNPs in the highest group, compared to 0% random SNPs.

2.3.5 Priors assigned through TFIDF and UMLS term expansion

In these experiments, terms chosen by TFIDF were expanded using UMLS.

Table 2.5 shows that combining TFIDF and UMLS term expansion, again fails to make a significant difference for breast cancer, but does for both prostate cancer and T2D. For T2D, 60% of true SNPs are placed in the highest prior group, compared to an expected proportion of 25%.



Breast cancer			Prostate cancer			Type 2 Diabetes		
Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c	Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c	Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c
0.00002	86%	63%	0.00002	92%	79%	0.00002	64%	34%
0.00014	13%	37%	0.00024	8%	15%	0.00024	36%	66%
0.00837	0%	0%	0.04302	0%	0%	0.01891	0%	0%
0.05980	0%	0%	0.36739	0%	0%	0.36757	0%	0%
0.13720	0%	0%	0.59215	0%	5%	0.59234	0%	0%
p value ^d	Average prior for random SNPs	Average prior for true SNPs	p value ^d	Average prior for random SNPs	Average prior for true SNPs	p value ^d	Average prior for random SNPs	Average prior for true SNPs
0.01	0.00013	6×10^{-5}	3×10^{-7}	9×10^{-5}	0.03	0.0004	0.00014	0.00016

Table 2.3: **Distributions of priors based on keywords generated through TFIDF** (a) The exact prior of each category was assigned based on (i) the observed distribution of random SNPs across the categories, and an (ii) assumed distribution and (iii) total number of truly associated SNPs. (b) The distribution was based on 598,805 SNPs included on the Illumina 610 quad genome-wide array. (c) True SNPs included SNPs previously robustly associated with breast cancer ($n = 19$), prostate cancer ($n = 39$), and Type 2 diabetes ($n = 35$). (d) The p -value is based on Chi-square statistics comparing the expected and observed distributions of 'truly' associated SNPs

2.4 Summary

We conducted an experiment aiming to evaluate if priors assigned to SNPs previously robustly associated with breast cancer, prostate cancer and type-2 diabetes, are higher than those assigned to random SNPs. We used several methods of generating keywords with which to calculate priors. For each method, the distribution of priors was found for random SNPs, and for known true positive SNPs. Comparing the two distributions with the Chi-square technique allows us to calculate a significance for each method. If a method is significant, then this means that ranking SNPs using priors calculated by this method, leads to significantly higher ranking of SNPs that are known to be associated with the disease. The results are summarised in Table 2.6.

For keywords generated by subjective assignment, random indexing and TFIDF, true SNPs were consistently assigned higher priors than random SNPs, directly implying that they would gain a higher ranking in GWAS if these priors were incorporated through BFD methods. We conclude that this methodology would be useful in GWAS, and may provide important cost savings by weighting SNPs that are more likely to be relevant to the disease higher than less relevant SNPs.

For keywords generated by UMLS expansion, true SNPs were only assigned higher priors than random SNPs for two of the three diseases considered. The reasons for this are unclear. Before rejecting term expansion, however, it is suggested that further experiments are required to both investigate and refine this technique.



Breast cancer			Prostate cancer			Type 2 Diabetes		
Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c	Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c	Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c
0.00002	99%	100%	0.00002	92%	79%	0.00002	64%	34%
0.00177	1%	0%	0.00024	8%	15%	0.00024	36%	66%
0.01032	0%	0%	0.01891	0%	0%	0.01891	0%	0%
0.22386	0%	0%	0.36757	0%	0%	0.36757	0%	0%
0.51782	0%	0%	0.59234	0%	5%	0.59234	0%	0%
p value ^d	Average prior for random SNPs	Average prior for true SNPs	p value ^d	Average prior for random SNPs	Average prior for true SNPs	p value ^d	Average prior for random SNPs	Average prior for true SNPs
0.55	1×10^{-4}	2×10^{-5}	3×10^{-7}	9×10^{-5}	0.030	0.0004	0.00014	0.00016

Table 2.4: **Distributions of priors based on keywords generated through UMLS expansion** (a) The exact prior of each category was assigned based on (i) the observed distribution of random SNPs across the categories, and an (ii) assumed distribution and (iii) total number of truly associated SNPs. (b) The distribution was based on 598,805 SNPs included on the Illumina 610 quad genome-wide array. (c) True SNPs included SNPs previously robustly associated with breast cancer ($n = 19$), prostate cancer ($n = 39$), and Type 2 diabetes ($n = 35$). (d) The p -value is based on Chi-square statistics comparing the expected and observed distributions of 'truly' associated SNPs

Breast cancer			Prostate cancer			Type 2 Diabetes		
Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c	Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c	Prior H_1^a	Expected SNP distribution ^b	Distribution for true SNPs ^c
0.00005	35%	26%	0.00004	34%	15%	0.00005	35%	14%
0.00007	0%	0%	0.00006	0%	0%	0.00007	0%	0%
0.00007	34%	37%	0.00007	53%	67%	0.00007	34%	26%
0.00031	5%	5%	0.00068	0%	0%	0.00031	5%	0%
0.00093	25%	32%	0.00170	13%	18%	0.00093	25%	60%
p value ^d	Average prior for random SNPs	Average prior for true SNPs	p value ^d	Average prior for random SNPs	Average prior for true SNPs	p value ^d	Average prior for random SNPs	Average prior for true SNPs
0.37	0.00029	0.00035	0.008	0.0002631	0.00035	3×10^{-6}	0.00029	0.00059

Table 2.5: **Distributions of priors based on keywords generated through TFIDF and UMLS expansion** (a) The exact prior of each category was assigned based on (i) the observed distribution of random SNPs across the categories, and an (ii) assumed distribution and (iii) total number of truly associated SNPs. (b) The distribution was based on 598,805 SNPs included on the Illumina 610 quad genome-wide array. (c) True SNPs included SNPs previously robustly associated with breast cancer ($n = 19$), prostate cancer ($n = 39$), and Type 2 diabetes ($n = 35$). (d) The p -value is based on Chi-square statistics comparing the expected and observed distributions of 'truly' associated SNPs



Method	<i>p</i> values		
	Breast cancer	Prostate cancer	Type 2 Diabetes
Subjective choice	0.004	4×10^{-05}	2×10^{-06}
Random indexing	0.04	0.001	4×10^{-08}
TFIDF	0.01	3×10^{-7}	0.0004
UMLS expansion	0.55	3×10^{-7}	0.0004
TFIDF with UMLS expansion	0.37	0.008	3×10^{-6}

Table 2.6: **Summary of keyword selection methods.** For each method, the distribution of priors was found for random SNPs, and for known true positive SNPs. These expected and observed distributions were compared using the Chi-square statistic, to give *p* values for every technique.



3. Qualitative Keyword evaluation

The previous section described the quantitative evaluation of prior assignment to SNPs by the LarKC GWAS service, for several techniques. The technique involved the generation of keywords associated with the disease being considered, and the use on those keywords to generate priors.

The quantitative evaluations gave an objective metric of keyword relevance (a p value). It is also interesting to see how relevant the keywords look to a domain expert: are subjectively relevant? This section presents a small set of results examining this. We do not pretend that these are more than very limited results. They do, however, give an indication of subjective relevance, and point to a further possible avenue of evaluation.

Keywords were generated by TFIDF over MEDLINE abstracts for three diseases (breast and prostate cancer and Type 2 diabetes). Keywords were generated by looking for terms commonly found in the same documents as the disease name. Two sets of documents were used:

- **Full MEDLINE** the whole of MEDLINE
- **MEDLINE subset** that subset of MEDLINE linked to genes by entries in Entrez Gene.

Keywords were presented to a domain expert, and the expert asked to select all keywords considered relevant to the disease in question. The results of this are shown in Table 3.1.

	Breast cancer		Prostate cancer		Type 2 Diabetes	
	MEDLINE subset	Full MED-LINE	MEDLINE subset	Full MED-LINE	MEDLINE subset	Full MED-LINE
Total terms	32	30	29	30	30	30
Relevant	13	6	16	14	16	11
Irrelevant	19	24	13	16	14	19
% Relevant	41	20	55	47	53	37
% Irrelevant	59	80	45	53	47	63

Table 3.1: Subjective relevance of keywords chosen by TFIDF

In four of six sets of keywords, most TFIDF terms were considered irrelevant. The exceptions were the MEDLINE subset for both prostate cancer and T2D. Clearly, TFIDF finds relevant terms as defined by a domain expert. These terms are, however, accompanied by a large proportion of noisy irrelevant keywords. It is interesting to note that the MEDLINE subset was the same as that used for the TFIDF experiments reported in Chapter 2, where TFIDF keywords gave significantly higher prior groups for true SNPs than random SNPs. Clearly, the noise of irrelevant keywords does not hinder the ability of the technique to assign high priors to true SNPs.



4. User evaluation

4.1 Introduction

As a final evaluation, we give a user-centric qualitative report of using the interface in practice.

4.2 Web service

LarKC has developed a web service for incorporating prior information from the literature into genome-wide association studies. The service offers an easy to use interface for users with limited experience in database handling and reasoning to initiate queries.

4.3 Advantages

The current implementation of the web service includes two separate parts, one for large scale queries of up to 15,000,000 SNPs (GWA service), and one for single SNP queries (SNP service). Extensive efforts have been made to provide an easy to use and accessible interface for less technical users. The first step of both services is the assignment of keywords for the endpoint of interest. This can be done both manually, as well as aided by various term expansion techniques:

- Random Indexing (RI)
- Term Frequency Inverse Document Frequency (TFIDF)
- Term expansion using UMLS

The GWA service is highly flexible as it does not require the users to provide their own data, but rather gives the results for all SNPs in the genome. The user can subsequently download the complete results and extract the prior for the relevant SNPs. This is an important advantage as many users would be expected to be reluctant to upload their own data. It is also possible to predefine the SNPs of interest, either through copy-pasting into a box, or by uploading a text file with SNP IDs. Typically a user would request priors for a set of SNPs included on their GWA chip. Once the SNPs and keywords of interest have been assigned the user will provide their email address and submit the query. Once the search is finished the user receives an email with a URL link for downloading the result, hence not requiring the user to create an account which increases the user-friendliness of the service. The downloaded priors can subsequently be used offline to calculate BFDP for re-ranking of SNPs. Overall, the GWA service is a remarkably streamlined interface that allows the non-technical users to perform complex queries over large data sets.

The single SNP service allows the user to assign single SNP literature queries. Based on key words that are assigned in the same fashion as in the GWA service, the result include all relevant abstracts with keywords highlighted in the text, as well as the relevant genes and prior assignment. Each abstract is accompanied by links to the related paper in Pubmed. Overall, the SNP service is an exceptionally convenient tool for literature searches and follow-up of initial results from genetic association studies.



4.4 Limitations and areas of improvement

While the user interfaces of both the GWA and SNP services are very streamlined without distracting frills, it would seem appropriate to include short texts next to the various search boxes with explanations of how things work. This could also be aided by longer documentations, perhaps with a small pop-up box. The final version of the GWA service should also provide scripts that allows the user to assign BFDP estimates in their own statistical package, e.g. for R and SAS.



5. Instrumentation applied on WP7b workflows

The previous evaluations of the GWAS prototype have focused only on the ability of the system to improve the SNP scoring/ranking. Through the instrumentation work documented in *LarKC deliverable D11.5, Evaluation of Instrumentation and monitoring tools and methods*[3] we can also evaluate the performance of the software. In this section we report on a number of experiments which expose the behavior of the whole platform, the workflows and plug-ins which constitute the WP7b prototype software as detailed in *LarKC deliverable D7b3.3b, Version 3 prototype*. We use the instrumented platform to evaluate the plugins and workflows in terms of execution time, CPU time and memory usage.

The GWAs workflow takes SPARQL queries, such as the following, as input.

```
PREFIX gwas: <http://www.gate.ac.uk/gwas#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT * WHERE {
  gwas:x rdf:type gwas:Experiment .
  gwas:x gwas:hasName "experiment1" .
  gwas:x gwas:hasKeywordGroup gwas:g1 .
  gwas:g1 gwas:hasKeyword "prostate" .
  gwas:g1 gwas:hasKeyword "cancer" .
  gwas:x gwas:searchInRif "false" .
  gwas:x gwas:useUMLS "false" .
  gwas:x gwas:searchMode "1" .
  gwas:x gwas:dateConstraint "20110412" .
  gwas:x gwas:hasSnpId "rs10007357" .
  gwas:x gwas:hasSnpId "rs10007361" .
  gwas:x gwas:hasSnpId "rs10007363" .
}
```

These queries specify the disease specific keywords used to score the specified SNPs.

For the purpose of evaluating the performance of the workflow and plugins under differing workloads we generated $O(10^4)$ different queries, using varying number (between 1 and 20) of SNPs. All the queries were then passed to the workflow which was hosted within version 2.5 of the LarKC platform (commit revision 1797), running on a 32-bit Windows machine with 4GB of memory.

5.1 Instrumentation Results

The instrumented platform collects a lot of data on the execution of workflows and plugins, and several views can be created over this data. In this section we focus upon high level views of execution time, CPU usage, memory usage and the number of threads allocated. A more thorough break down and interpretation of the collected data can be found in [3].

The CPU behaviour during execution of the workflow can be seen in Figure 5.1. Things to note are that...

- the platform CPU time denotes the total CPU time spent by all processes comprising the LarKC platform.

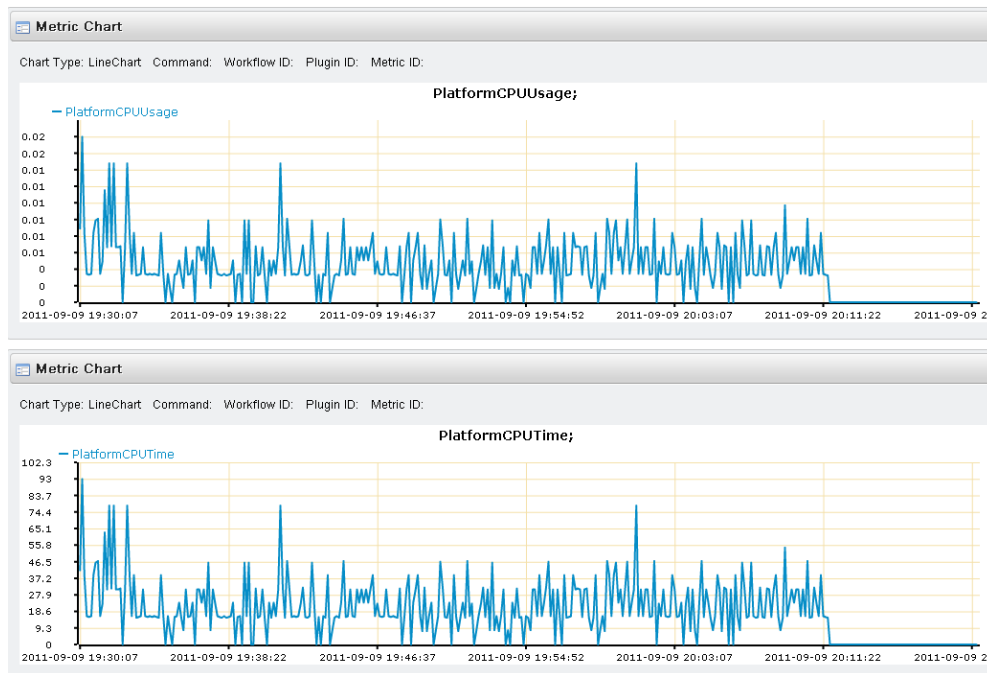


Figure 5.1: CPU Usage and and Time

- The platform CPU usage shows the degree of parallelism within the application. Notice that maximum CPU usage appears during initialization of the platform and hence the running of the workflow does not involve excessive CPU usage.

In comparison the memory allocation during the running of the workflow can be seen in in Figure 5.2. One issue with these evaluations are that they cover a (relatively) long period of time and many passes through the workflow. While this allows us to see long term trends it makes it difficult to evaluate the performance for a single query. We can, however, examine a narrower time interval.

Figure 5.3 shows how the size of the query (in characters, which relates directly to the complexity of the query) effects the response time and CPU usage. The Total Response Time represents the time elapsed from when the query was sent to the platform until the final result is provided to the user. It would appear from these results that the query time is extremely variable and appears to have little if any relationship with the size of the query. This could be due to a number of reasons including disk access and caching. Figure 5.3 also shows the thread block count, which shows how often the thread in which the plugin is running enters the blocked state. The more time the thread spends blocked the longer the execution time. Thread blocking may well explain the erratic query response time but further, more detailed, investigations are required in order to determine the exact reasons for these results.



Figure 5.2: Amount of Free, Allocated and Memory Used

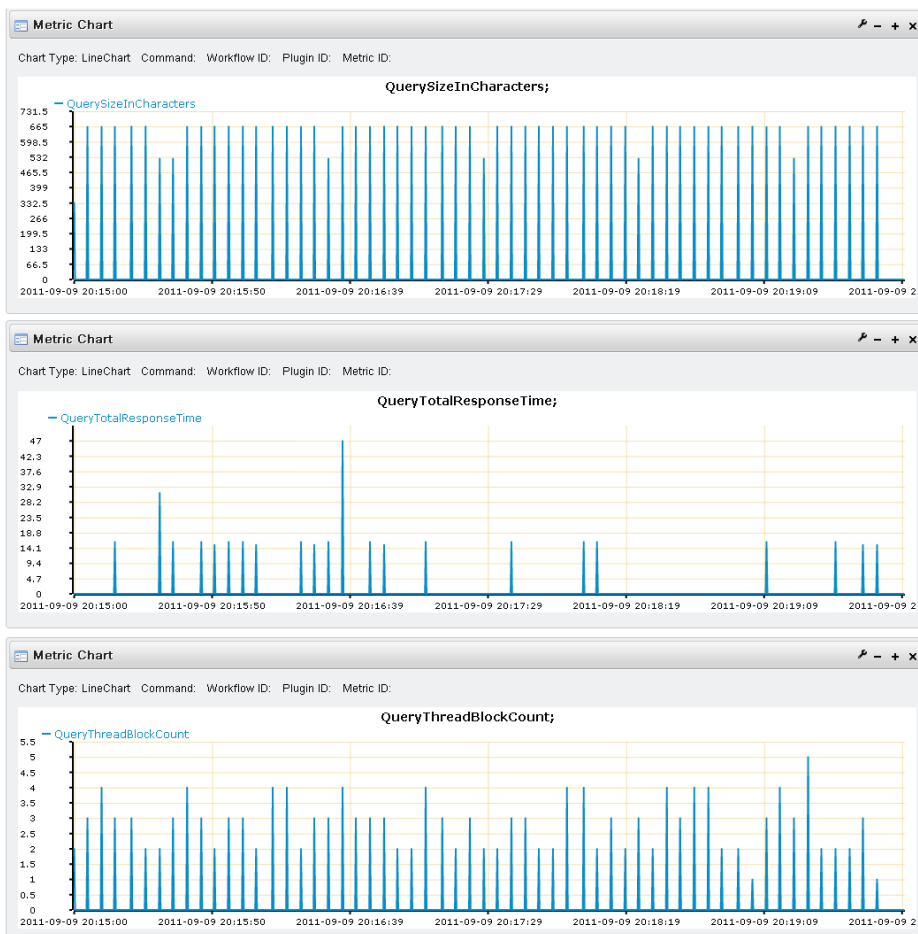


Figure 5.3: Query Level Instrumentation Results



6. Conclusion

We have presented an evaluation of the LarKC GWAS service. The service finds prior knowledge about gene-disease associations in the scientific literature and biomedical knowledge resources. A prior probability is associated to SNPs according to the occurrence of keywords and concepts in this prior knowledge. These prior probabilities can then be combined with GWAS experimental data in a Bayesian model in order to predict gene-disease associations.

The service is data intensive, with very little user interface. It has therefore been felt that a quantitative evaluation of the service is most relevant. We have presented such an evaluation, comparing the priors assigned to SNPs with known disease associations to random SNPs. It would be expected that a successful GWAS service would give consistently higher priors to such true SNPs when compared to random SNPs. We examined three diseases, and several ways of generating keywords. Subjective keyword assignment, random indexing and TFIDF all gave good results, with significantly higher priors for true SNPs. UMLS expansion of keywords did not always give significantly higher priors for true SNPs, and needs further evaluation.

We have also included a subjective evaluation of the user interface by end users. While the interface is by its nature limited, end users consider it sufficient and fit for purpose.



REFERENCES

- [1] A.Roberts, M. Greenwood, D. Damljanovic, H. Cunningham, T. Heitz, I. Roberts, Y. Li, M. Johannson, and J. McKay. D7b.3.1b version 1 prototype. Technical report, LarKC project deliverable, 2009.
- [2] Niraj Aswani, Mark Greenwood, Mattias Johansson, Angus Roberts, James McKay, Jon Wakefield, Valentin Tablan, Ian Roberts, Hamish Cunningham, and Paul Brennan. D7b.3.2b version 2 prototype. Technical report, LarKC project deliverable, 2010.
- [3] Raluca Brehar, Ioan Toma, Silviu Bota, Ionel Giosan, Mihai Negru, Andrei Vatavu, and Mihai Chezan. Evaluation of instrumentation and monitoring tools and methods. Technical Report D11.5, LarKC Project Deliverable, 2011.
- [4] Mark A. Greenwood, Mattias Johansson, Niraj Aswani, Angus Roberts, James McKay, Jon Wakefield, Valentin Tablan, Ian Roberts, Hamish Cunningham, and Paul Brennan. D7b.3.3b version 3 prototype. Technical report, LarKC project deliverable, 2010.
- [5] Mattias Johansson, Niraj Aswani, Mark Greenwood, Angus Roberts, James McKay, Jon Wakefield, Valentin Tablan, Ian Roberts, Hamish Cunningham, and Paul Brennan. D7b.3.2a version 2 iteration report. Technical report, LarKC project deliverable, 2010.
- [6] A. Roberts, H. Cunningham, and A. Funk. D7b.1.1b iteration evaluation methodology and report template. Technical report, LarKC project deliverable, 2008.
- [7] A. Roberts, K. Straif, J. McKay, M. Stetter, and H. Cunningham. D7b.1.1a requirements summary and data repository. Technical report, LarKC project deliverable, 2008.
- [8] Angus Roberts, Mark Greenwood, Danica Damljanovic, Hamish Cunningham, Mattias Johansson, and James McKay. D7b.3.1a version 1 iteration report. Technical report, LarKC project deliverable, 2009.