Genetics and Molecular Biology, 32, 3, 634-639 (2009)
Copyright © 2009, Sociedade Brasileira de Genética. Printed in Brazil
www.sbg.org.br

Research Article

D-MaPs - DNA-microarray projects: Web-based software for multi-platform microarray analysis

Marcelo F. Carazzolle, Taís S. Herig, Ana C. Deckmann and Gonçalo A.G. Pereira

Laboratório de Genômica e Expressão, Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, SP, Brazil.

Abstract

The web application D-Maps provides a user-friendly interface to researchers performing studies based on microarrays. The program was developed to manage and process one- or two-color microarray data obtained from several platforms (currently, GeneTAC, ScanArray, CodeLink, NimbleGen and Affymetrix). Despite the availability of many algorithms and many software programs designed to perform microarray analysis on the internet, these usually require sophisticated knowledge of mathematics, statistics and computation. D-maps was developed to overcome the requirement of high performance computers or programming experience. D-Maps performs raw data processing, normalization and statistical analysis, allowing access to the analyzed data in text or graphical format. An original feature presented by D-Maps is GEO (Gene Expression Omnibus) submission format service. The D-MaPs application was already used for analysis of oligonucleotide microarrays and PCR-spotted arrays (one- and two-color, laser and light scanner). In conclusion, D-Maps is a valuable tool for microarray research community, especially in the case of groups without a bioinformatic core.

Key words: microarray, web service, software, affymetrix and nimblegen.

Received: December 26, 2008; Accepted: May 6, 2009.

Introduction

Biological approaches based on DNA microarrays are becoming instrumental to gain insights into gene expression from a genomic perspective. Since the technique supports the representation of the whole set of genes or genomic features on a single glass slide, virtually any cellular response can be investigated in detail. Moreover, microarray data analysis allows researchers to assess not only gene expression or regulation, but also gives support for characterization of gene functions and interactions.

However, despite the fact that microarrays are currently one of the most widespread techniques for functional studies in the fields of biology and medicine, it is extremely difficult to perform comparisons of the available data. First, differences in experimental design and microarray protocols must be considered when comparing gene expression data. The other problematic steps are mainly computational issues: data preprocessing, including data normalization, and statistical analysis for identification of differentially expressed genes. The preprocessing step is specific for each platform and aims to reduce technical artifacts with minor alterations to biological information. The statistical analysis summarizes the biological and technical replicates data by generating one value for the expression data (fold-change) and one value for the statistical result (p-value).

In an attempt to systematize analysis from the initial steps, by including raw data preprocessing, we describe here a web-based system - D-MaPs (DNA-Microarray Projects) - developed to manage and process microarray data obtained from different experimental designs (one- and two-color fluorescent labeling), equipment (light and laser scanners) and platforms (PCR-product spotted arrays and oligonucleotide arrays) (Schena *et al.*, 1995; Nuwaysir *et al.* 2002). The system has been implemented to manage GeneTAC, ScanArray, CodeLink, NimbleGen and Affymetrix microarray platforms.

D-MaPs features a simple interface to allow researchers unfamiliar with statistics and graphical development to easily obtain their results. The user can access the system using the version installed on a server located in our bioinformatics laboratory. D-Maps system also provides the output data in GEO (Gene Expression Omnibus) submission format, commonly requested by many periodicals for articles based on microarray data.

Send correspondence to Marcelo F. Carazzolle. Laboratório de Genômica e Expressão, Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, Caixa Postal 6109, 13083-970 Campinas, SP, Brazil. E-mail: mcarazzo@lge.ibi.unicamp.br.

Methods

The D-MaPs application was written in Perl and R computer languages, supports entry of various file formats and contains implementations of diverse Bioconductor packages (Gentleman *et al.*, 2004) for normalization and statistical analysis. It uses standard open source modules such as CGI.pm and DBI.pm. The system requires the Linux operating system with a Web server (*e.g.*, Apache) and MySQL server installed. The system runs on a high performance server with large amount of memory (16 Gb of ram). The main advantage of this architecture is that the server performs the majority of the processing work, which makes it unnecessary for external users to have high performance computers.

With exception of the database management and dynamic HTML pages generation, all calculations are performed in R functions of the Bioconductor packages including graphics generation. The R scripts are dynamically generated while the user navigates on the interface.

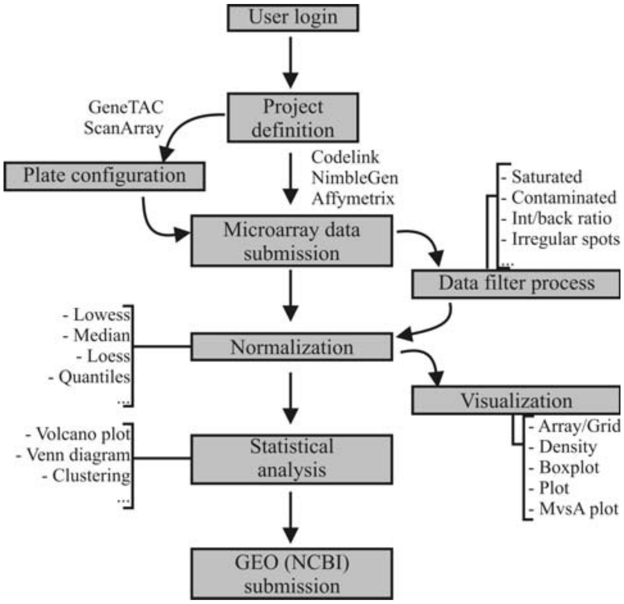


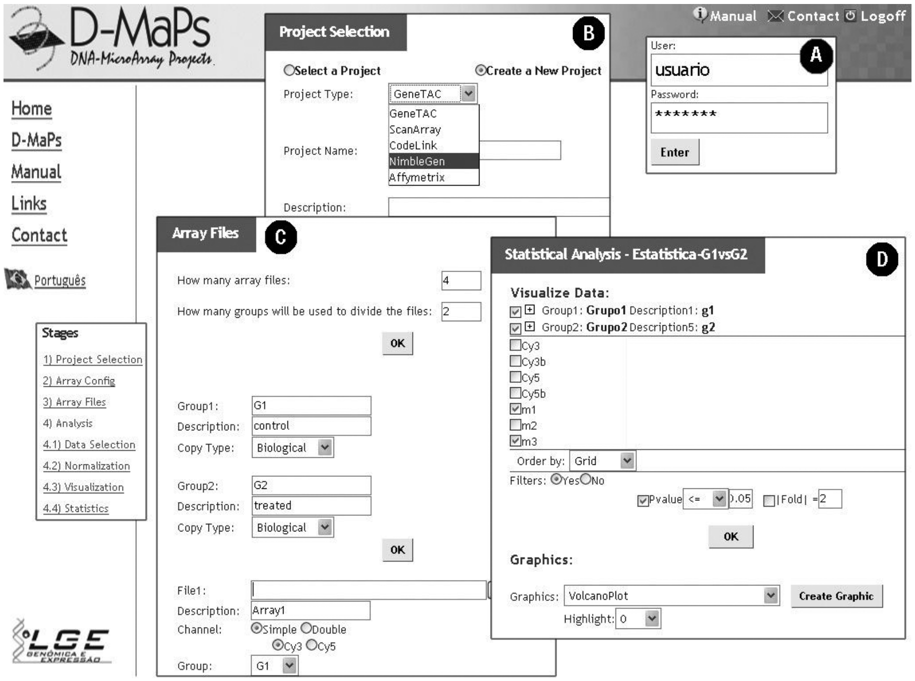
Figure 1 - The D-maps analysis workflow. The web interface guides the users through these steps.

Program Description

For each user D-maps program provides a separate workspace, called projects, for storing and analyzing microarray data. The web interface guides the user through further steps, such as project definition, upload data, data preprocessing and statistical analysis, necessary to perform the analysis of microarray data. During each step the user has access of all analyzed data in text or graphical format. Figure 1 shows the main features of the system.

Project description and data submission

First of all, the user creates a login/password (Figure 2A). Then, he is prompted to create a project and provide some details about it, including type of scanner (platform)



User: usuario
Password: *******

Project Selection

○Select a Project ●Create a New Project
Project Type: GeneTAC
GeneTAC
ScanArray
CodeLink
NimbleGen
Affymetrix
Project Name:
Description:

Stages
1) Project Selection
2) Array Config
3) Array Files
4) Analysis
4.1) Data Selection
4.2) Normalization
4.3) Visualization
4.4) Statistics

Array Files

How many array files: 4
How many groups will be used to divide the files: 2

Group1: G1
Description: control
Copy Type: Biological

Group2: G2
Description: treated
Copy Type: Biological

File1:
Description: Array1
Channel: ●Simple ○Double
●Cy3 ○Cy5
Group: G1

Statistical Analysis - Estatistica-G1vsG2

Visualize Data:
☑ ⊞ Group1: **Grupo1** Description1: **g1**
☑ ⊞ Group2: **Grupo2** Description5: **g2**
☐Cy3
☐Cy3b
☐Cy5
☐Cy5b
☑m1
☐m2
☑m3
Order by: Grid
Filters: ●Yes○No
☑Pvalue <=).05 ☐|Fold| =2

Graphics:
Graphics: VolcanoPlot
Highlight: 0

Figure 2 - Some D-maps Web interfaces: (A) User and password control; (B) Project description; (C) Array details and upload data - in this step the user defines the groups and provides details about each array file (One or two channel, dye swap and experimental replicates); (D) Statistical analysis - the user defines the threshold values for fold-change and p-value to obtain the differential expressed genes.

used for image acquisition (selection of 1 of 5 platforms currently supported by D-Maps: GeneTAC, ScanArray, CodeLink, NimbleGen and Affymetrix) and a basic description. The project definition interface is shown in Figure 2B.

The next stage is designed for the submission of raw data files (Figure 2C). The files can be divided into groups according the experimental design (*e.g.*, one group is formed by files representing data from experimental samples and other group from control samples). In the case of projects based on GeneTAC or ScanArray platforms, it is also necessary to define whether microarrays are one- or two-color (*i.e.*, single or dual channel images), and if the file represents the dye-swap replicates or not. For Affymetrix, Codelink and NimbleGen platforms, the raw data files are defined as single channel. Also, the file formats are different according to the platform used: CSV file for GeneTac and ScanArray platforms, TXT file for Codelink, XYS and NDF files for NimbleGen and CEL files for Affymetrix.

Data preprocessing

Data preprocessing is an essential step in the analysis of microarray data. Basically, it consists of filtering the data using scanner flag information, background correction, normalization intra- and between-slides calculations. The difficulty in this step is that raw data generated by each platform is flagged by a specific set of identifiers according methods and parameters that are platform specific. D-maps can perform this step using the best set of methods and parameters for each platform, but the user can change this information depending on particularities of the data. The graphical tools are important to resume the information contained in the microarray data during the preprocessing steps. Figure 3 shows some graphics displayed by D-Maps.

Flag determination. After submission of raw data files, an automatic (adjustable) assignment is made based on the quality of intensity signals (saturated, contaminated, irregular), excluding artifactual data prior to further analysis. As mentioned above, the flags are platform dependent.

Background correction. Usually, the spot intensities are corrected for background intensities by subtraction of the latter. For the Affymetrix platform, the specific background subtraction methods are available to users: RMA and MAS algorithms (Eschrich and Hoerter, 2007). For the Codelink platform, the Subtract method (Diez *et al.*, 2007) is implemented.

Data normalization. A large number of normalization tools are available in this step for one- or two-color microarrays. All normalization methodologies perform the log transformation (base 2) of fluorescence intensities prior to the analysis. In the case of two-color microarrays (available for GeneTac and ScanArray platform), the log ratio of intensities (Log2 Cy5/Cy3) is calculated.

Two-color microarrays. For two-color experiments the system performs both intra- and between-slides calculations (Irizarry *et al.*, 2003; Yang *et al.*, 2002). The intra-

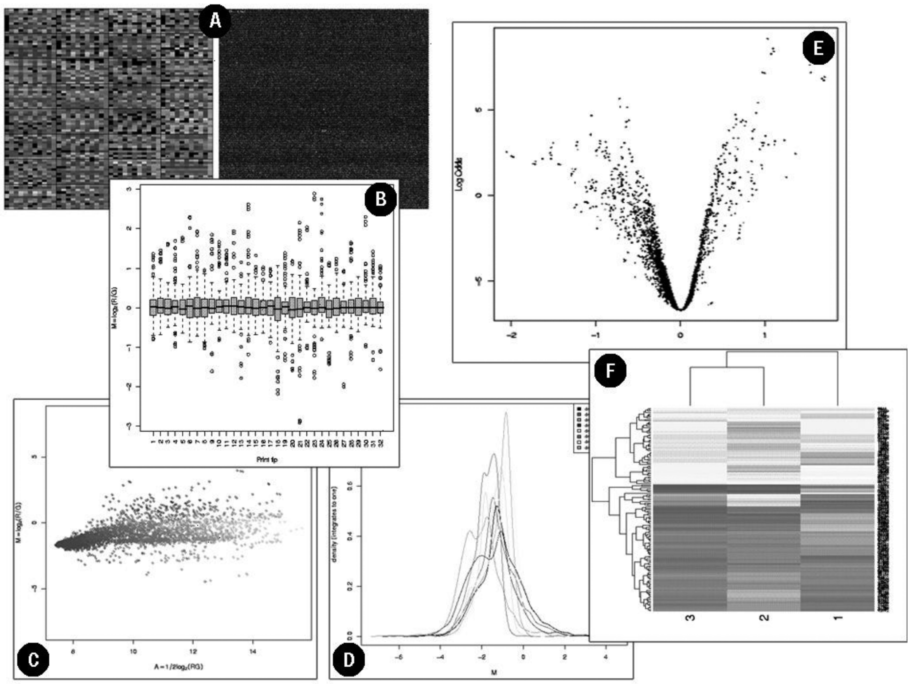


Figure 3 - Graphics visualizing some steps during the analysis workflow: (A) spatial plot showing the spot intensities; (B) Box plot showing the grid variation in one array; (C) M-A plot summarizing the spot intensities in a two channel experiment; (D) Density plot summarizing the distribution of intensities between arrays; (E) Volcano plot showing the differential expressed genes; (E) Cluster analysis grouping the genes with similar expression pattern.

slide normalization process is necessary to remove such specific variations, *e.g.* differences in labeling efficiency of the two fluorescent dyes. D-Maps allows the user to choose among Lowess-scaled print-tip (Yang *et al.*, 2002), Lowess print-tip (Yang *et al.*, 2002), Global Lowess (Yang *et al.*, 2002) and Median (Quackenbush, 2002) methods to perform the intra-slide normalization. After that, all normalized log-ratios will be centered around zero. Since distinct slides present non-homogeneous variations in their log-ratio values (the so-called "scale effect"), it must be corrected using between-slides normalization methods. D-maps performs between-slides normalization using Quantile method (Bengtsson and Bengtsson, 2006).

One-color microarrays. In the case of Codelink platform, just the Cyclic Lowess method is implemented (Wu *et al.*, 2005; Diez *et al.*, 2007) for data normalization. For Nimblegen platform, the Quantile normalization is implemented. Finally, for the Affymetrix Genechip platform, the normalization methods were implemented using RMA and MAS algorithms (loess, constant, contrasts, invariantSet, qspline, quantiles and quantiles.robust).

Statistical analysis applied to detection of differentially expressed genes

After normalized, data is statistically analyzed both for reproducibility of replicated values and for significance of the differences among conditions to be compared. To perform this step, D-Maps has implemented the Limma package (Smyth, 2004) from Bioconductor. Limma uses linear models to analyze the microarray experiments (eBayes-moderated t-test). The statistical analysis summarizes the biological and technical replicates data by generating two values for each probe: (i) fold-change value: summarizes the intensity of each gene by calculating the average of normalized values in each experimental replicate; (ii) p-value: summarizes the coherence between the normalized values in each experimental replicate. Figure 2D shows the statistical analysis interface, in which the user is prompted to define the threshold values for fold-change and p-value. This set of parameters generates the set of differentially expressed genes, which are displayed graphically (Volcano and clustering plot, shown in Figure 3E and 3F, respectively) and in plain text format (txt files).

Results and Discussion

In recent years, Bioconductor has become the reference tool for the analysis of microarray data because it is based on up-to-date algorithms. However, for scientists without programming experience, the command line usage of Bioconductor is too difficult. Therefore, many analysis tools with a graphical user interface and powerful computing servers have been developed, including web-based tools like Webarray (Xia *et al.*, 2005), MIDAW (Romualdi *et al.*, 2005), MAGMA (Rehrauer *et al.*, 2007) or CarmaWEB (Rainer *et al.*, 2006).

In most cases, however, there are limitations concerning platforms and type of data formats supported by these software programs. For instance, data from two-color microarrays requires the external manipulation before the input step. Since the raw data files derived from the image analysis software are usually large and difficult to handle, especially for inexperienced users, researchers working with two-color microarray data normally have to navigate several websites and transfer the data between the servers to complete their analysis. With this in mind, D-Maps software was designed and implemented. Table 1 presents a comparative view of the main features of Webarray,

Table 1 - Some surveyed systems and their characteristics.

Characteristics	WebArray	MIDAW	MAGMA	CarmaWeb	D-maps
Web interface	X	X	X	X	X
User and project management	X				X
One channel microarray data	X			X	X
Two channel microarray data	X	X	X	X	X
Preprocessing data analysis	X	X	X	X	X
Detection of differentially expressed	X	X	X	X	X
Visualization tools	X	X	X	X	X
Cluster analysis		X		X	X
Submission at GEO database (NCBI)					X
Microrray platforms supported:					
Affymetrix	X			X	X
Codelink			X		X
GeneTac		X		X	X
Nimblegen					X
ScanArray		X		X	X

MIDAW, MAGMA, CarmaWEB and D-Maps software programs.

An original feature presented by D-Maps software is GEO (Gene Expression Omnibus) submission format service. Nowadays, scientific articles based on microarray data must be accompanied by submission of both raw and processed data to public databases, as NCBI/GEO Database (Barrett *et al.*, 2007). This process requires data to follow some format standards, including experimental design and technical requirements of platforms. D-Maps prompts users to give all necessary information during raw data submission, therefore output files containing GEO formats are automatically generated.

The D-MaPs application has already been used for analysis of oligonucleotide microarrays (one-color, laser scanner) (Deckmann AC, Rocco SA, Marin RM, Rolim Filho LA, Carazzolle MF, Barau JG, Pereira GAG and Franchini KG, unpublished data) and PCR-spotted arrays (one- and two-color, laser and light scanner) (Rincones *et al.*, 2008; Rubin MS, Deckmann MF, Carazzolle MF, Parizzi LP, Pereira GAG and Annichino-Bizzacchi JM, unpublished data; Lepikson-Neto J, Deckmann AC, Carazzolle MF and Pereira GAG, unpublished data). In the first study, D-Maps performed the pre-processing of raw data obtained from Codelink bioarrays according Codelink package of Bioconductor repository: inter-slide normalization by Lowess method and differential expression by eBayes-moderated t-test for microarrays. In later studies, raw data obtained from distinct scanners were analyzed by D-Maps according to the Aroma and Limma package of Bioconductor repository: intra-slide normalization (Lowess or median), inter-slide normalization (Quantile) and differential expression by eBayes-moderated t-test for microarrays. Other studies are currently in progress, for which GEO submission support service has been implemented.

Conclusions

In this study we describe a web-based system, D-MaPs (DNA-Microarray Projects), developed to manage and process microarray data obtained from distinct experimental designs (one- and two-color fluorescent labeling) and platforms (GeneTAC, ScanArray, CodeLink, NimbleGen and Affymetrix). D-MaPs features a simple interface to allow researchers who are unfamiliar with computation and statistics to easily obtain results of differential gene expression in graphical and plain text formats. The D-Maps system also provides output data in GEO (Gene Expression Omnibus) submission format. In conclusion, D-Maps is a valuable tool for microarray research community, especially in the case of groups without a bioinformatic core.

Acknowledgments

The work described in this paper was partially financed by FAPESP, CNPq and CAPES. We acknowledge the researchers that were or are end-users of our system and assisted us in improving the functionalities, interfaces and the general quality of D-Maps.

References

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M and Edgar R (2007) NCBI GEO: Mining tens of millions of expression profiles - Database and tools update. Nucleic Acids Res 35:D760.

Bengtsson A and Bengtsson H (2006) Microarray image analysis: Background estimation using quantile and morphological filters. BMC Bioinform 7:96.

Diez D, Alvarez R and Dopazo A (2007) Codelink: An R package for analysis of GE healthcare gene expression bioarrays. Bioinformatics 23:1168-1169.

Eschrich SA and Hoerter AM (2007) Libaffy: Software for processing Affymetrix(R) GeneChip(R) data. Bioinformatics 23:1562-1564.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al.* (2004) Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol 5:R80.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249-264.

Nuwaysir EF, Huang W, Albert TF, Singh J, Nuwaysir K, Pitas A, Richmond T, Gorski T, Berg JP, Ballin J, *et al.* (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res 12:1749-1755.

Quackenbush (2002) Microarray data normalization and transformation. Nat Genet 32:496-501.

Rainer J, Sanchez-Cabo F, Stocker G, Sturn A and Trajanoski Z (2006) CARMAweb: Comprehensive R- and bioconductorbased web service for microarray data analysis. Nucleic Acids Res 34:W498.

Rehrauer H, Zoller S and Schlapbach R (2007) MAGMA: Analysis of two-channel microarrays made easy. Nucleic Acids Res 35:W86.

Rincones J, Scarpari LM, Carazzolle MF, Mondego JMC, Formighieri EF, Barau JG, Costa GGL, Carraro DM, Brentani HP, Vilas-Boas LA, *et al.* (2008) Differential gene expression between the biotrophic-like and saprotrophic mycelia of the witches' broom pathogen *Moniliophthora perniciosa*. Molecular Plant-Microbe Interactions 21:891.

Romualdi C, Vitulo N, Favero MD and Lanfranchi G (2005) MIDAW: A web tool for statistical analysis of microarray data. Nucleic Acids Res 33:W644.

Schena M, Shalon D, Davis RW and Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467-470.

Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:3.

Wu W, Dave N, Tseng GC, Richards T, Xing EP and Kaminski N (2005) Comparison of normalization methods for CodeLink Bioarray data. BMC Bioinform 6:309.

Xia X, McClelland M and Wang Y (2005) WebArray: An online platform for microarray data analysis. BMC Bioinform 6:306.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP (2002) Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30:e15.

Internet Resources

Affymetrix Genechip microarray platform, http://www.affymetrix.com (October 15, 2008).

CodeLink microarray platform, http://www4.gelifesciences.com (October 15, 2008).

GeneTAC microarray platform, http://www.genomicsolutions.com (October 15, 2008).

NimbleGen microarray platform, http://www.nimblegen.com/ (October 15, 2008).

Perl5 library for writing World Wide Web CGI scripts, http://stein.cshl.org/WWW/software/CGI (October 15, 2008).

ScanArray microarray platform, http://las.perkinelmer.com (October 15, 2008).

Standard database interface module for Perl, http://dbi.perl.org (October 15, 2008).

The D-maps web application server, http://www.lge.ibi.unicamp.br/d-maps (October 15, 2008).

Oligonucleotide Arrays: A package to analyze oligonucleotide arrays (expression/SNP/tiling/exon) at probe-level, http://www.bioconductor.org/packages/2.0/bioc/html/oligo.html (October 15, 2008).

R programming language, http://www.r-project.org (October 15, 2008).

Perl programming language, http://www.perl.org (October 15, 2008).

The open source database, http://www.mysql.com (October 15, 2008).

Apache software foundation, http://www.apache.org (October 15, 2008).

Guest Editor: José Carlos Merino Mombach