Preclinical Tests for Cerebral Stroke

M.F. Zini
Department of Computer Science
University of Pisa

N. Pisanti
Department of Computer Science
University of Pisa
&
ERABLE Team, INRIA

E. Biasci, A. Podda, V. Mey, F. Piras,
G.L. L'Abbate, S. Marini, D. Fratta
Galileo Research S.r.l.

S. Bonaretti
Galileo Research S.r.l.

S. Trasciatti
Galileo Research S.r.l.

***Abstract*—Stroke is the second single highest cause of death in Europe. The low reliability of animal models in replicating the human disease is one of the most serious problems in the field of medical and pharmaceutical research about stroke. The standard models for the study of ischemic stroke are often poorly predictive as they simulate only partially the human disease. This work aims at investigating animal models with diseases typically associated with the onset of stroke in human patients.**
We have designed and realised a knowledge base for collecting, elaborating, and extracting analytical results of genomic, proteomic, biochemical, morphological investigations from animal models of cerebral stroke. Data analysis techniques are tailored to make the data available for processing and correlation, in order to increase the predictive value of the preclinical data, to perform biosimulation studies, and to support both academic and industrial research in the area of cerebral stroke therapy. A first statistical analysis of the retrieved information leads to the validation of our animal models and suggests a predictive and translational value for parameters related to a specific model. In particular, concerning gene expression data, we have applied a data analysis pipeline that initially takes into account an initial set of 64,000 genes and brings down the focus on a few tens of them.

I. Introduction

Stroke is the second single highest cause of death in Europe [1] and more in general in the developed countries, the third in UK, and the fourth in the US. Up to 80 percent of strokes could be prevented. The low reliability of animal models in replicating the human disease is one of the most serious problems in the field of medical and pharmaceutical research about stroke. The standard models for the study of ischemic stroke are often poorly predictive as they simulate only partially the human disease. This work aims at investigating animal models with diseases typically associated with the onset of stroke in human patients. The purpose of this study is the evaluation of the evolution of different data types (with a special focus on gene expression data) among several strains of rats subject to different treatments that stimulate stroke, and some possible stroke consequences. This study is mainly intended to assess whether these new animal models are more consistent and predictive of the human condition. This increased knowledge could also reduce the number of animals used in these kinds of experimentations.

We have designed and realised a knowledge base (a data base) for collecting, elaborating, and extracting analytical results of genomic, proteomic, biochemical, morphological investigations from animal models of cerebral stroke. Data analysis techniques are tailored to make the data available for processing and correlation, in order to increase the predictive value of the preclinical data, to perform biosimulation studies, and to support research in the area of cerebral stroke therapy. A first statistical analysis of the retrieved information leads to the validation of our animal models and suggests a predictive and translational value for parameters related to a specific model. In particular, concerning gene expression data, we have applied a data analysis pipeline that initially takes into account an initial set of 64,000 genes and brings down the focus on a few tens of them.

This paper is organised as follows. Section II introduces the data used (strains and treatments), the group models obtained for the comparative analysis, and the data types acquired for each group. Section III describes the method applied for the data collection and information retrieval, and the data analysis method for the gene expression data type, whose results are the focus of this paper. Section IV describes the results of such analysis and, finally, Section V concludes the paper.

II. Data and Models

The animal models were obtained with techniques of transient or permanent occlusion of the middle cerebral arteries, adapted for different types of animal groups.

A. Data

Three different strains of rat have been used for this preclinical study:

SD	Sprague Dawley Rat.
ZL	Zucker Lean Rat. (Zucker strain, healthy control).
ZDF	Zucker Diabetic Fatty Rat. (Zucker strain, diabetic/hypertensive/obese rats).

For all such strains, we have performed three different treatments:

S	Sham operated.
T	Transient occlusion.
P	Permanent occlusion.

Sham surgery is a faked operation where the control groups have received the same surgical procedure as the others, but without the occlusion procedure that induces ischemia. This allows to distinguish scientific data that reflect the effects of the experiment itself from that which is a consequence of the surgery.
The transient occlusion consists in applying an occlusion in

the middle cerebral arteries within a surgery, and then remove it after two hours, while the permanent occlusion is never removed. The distinction of these two treatments is meant to investigate the effects of *reperfusion*. Indeed, the blood supply after lack of oxygen can severely damage tissues ([2]) and can actually cause more damage than the actual ischemic event.

Finally, the data we aim to collect for our preclinical study is the following:

- Haematological data
- Blood chemistry data
- Gene expression data (MicroArray Analysis, NGS miRNA)
- Physiological data (body/organs weight and blood pressure)
- Enzyme activity data
- Histological data (optical microscopy)
- Receptor Binding data
- Neurological data (through behavioural observation and scores)
- Protein expression data (matrix expression)
- Mitochondrial damage data (SEM: Scanning Electron Microscopy)
- Proteomics data (Ms spectrometry, 2D electrophoresis)

B. The model

In Section II-A we described the strains and their possible treatments. This led to the following groups for our preclinical analysis:

SDS	*Sham operated SD.*
SDP	*SD with Permanent occlusion.*
ZLS	*Sham operated ZL.*
ZLT	*ZL with Transient occlusion.*
ZDFS	*Sham operated ZDF.*
ZDFT	*ZDF with Transient occlusion.*

As explained above, the sham-operated groups SDS, ZLS, and ZDFS are the control groups. All the others are those on which either transient or permanent ischemia is investigated. In the preclinical model, the ZDF strain has been chosen to test the effect of the treatments on animals that simulate the conditions of patients that are at high risk for ischemia. Unfortunately, no data could be collected for permanent occlusion of this group as ZDF rats did not survive the treatment, and this is why there is no *ZDFP* group. For the same reason, we did not perform permanent occlusion treatment on ZL strain, and hence we do not have a *ZLP* group, because the latter would have served as a control group for the missing *ZDFP*. In fact, we collected data to investigate permanent occlusion only for SD strain.

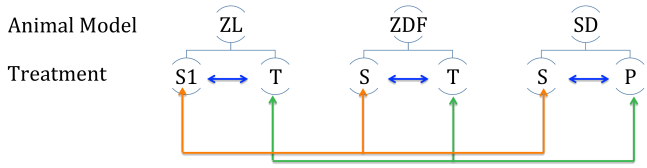

Fig. 1. Groups comparisons of interest

III. Method

A. Data Base and Information retrieval

As we have seen in last section, data exhibits strong heterogeneity. Moreover, the genomic, proteomic, biochemical and morphological investigations provide information of different multiplicity and with different number of attributes. We have designed and realised a knowledge base that collects all this data types (and that can also host possible new kind of data) in an updated, detailed and non-redundant way. The chosen DBMS software is PostgreSQL, and the script language used for reading and inserting data is python. SQL allows to extract data for each performed analysis, or aggregated data for type of analysis or group or specimen, and combinations. The database, by means of logical views to integrate information, can directly export data to the analysis workflow even for the kind of results with multiple answers per sample (microarrays, MALDI MS/MS, NGS).

Figure 1 shows, for each animal model and treatment, and hence for all groups, what are those for which we are interested in crossing the information by comparing data. We compare: (i) each Sham operated animal model with its corresponding treated group (blue arrows), (ii) all Sham operated animal models among themselves (orange), and (iii) all treated models (green). Finally, a downstream comparative analysis will have to consider them all.

B. Data Analysis

In this section we show a data analysis we performed on the specific data type of gene expression.

For high-throughput data analysis we used methods from *R-Bioconductor* project ([3]), namely:

- *limma:* differential gene expression analysis for microarray data based on linear models and moderated T-statistics ([4],[13],[6]).
- *EdgeR* and *DeSeq:* differential expression analysis for count data based on negative binomial distribution ([8],[9],[5]).

For the Microarray analysis of the transcriptome, the following workflow was performed:

- Microarray scansion with Agilent one-color SurePrint G3 Rat GE 8x60K Microarray Kit.
- The generated raw data (a 64K records matrix per sample) were collected with Feature Extraction Software (Agilent proprietary software).

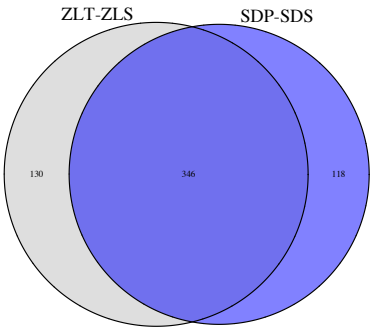

Fig. 2. The blue (right) circle represents the set of genes differentially expressed in the two condition (Sham and treatment) for the SD Strain, while the grey (left) circle is that of the ZL strain in its two conditions. Numbers (and circles' size) highlight the number of elements in the intersection and in the sets differences .

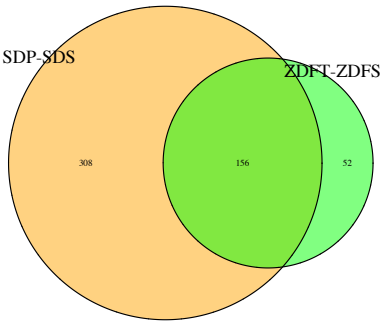

Fig. 3. The right circle represents the set of genes differentially expressed in the two condition (Sham and treatment) for the ZDF Strain, while the left circle is that of the SD strain in its two conditions.

- Array quality was monitored using Agilent Feature Extraction Software QC Report to evaluate Microarray Performance for every sample. Samples whose QC report did not pass the quality control threshold were excluded.
- Data from all samples was merged into a single gene expression matrix collecting all samples that survived previous steps. Each single group now corresponds to a set of (contiguous) columns of this new matrix.
- Probes filtering: control probes and too low expressed genes were removed.
- Normalization was performed with quantile normalization between single groups arrays.
- Differential expression analyses were performed using two different techniques:
  - Empirical Bayes-moderated F-statistics with the limma Bioconductor package.
  - 1 way Anova with SNK post hoc test with the GeneSpring Agilent software.
- The two statistics gave comparable results. limma has been choosen to perform further analysis.
- Limma multi-group analysis has been performed on data. We obtained lists of differently expressed genes using threshold parameters: $P - adjusted < 0.01/0.05$ and $FoldChange > 2$.
- Unsupervised analysis (clustering) with different algorithms was performed on differentially expressed (DE) genes, to see if genes differently expressed for one contrast were giving good clustering also for other classes. Clustering was performed by Amic@ server, with K-means method and Pearson Correlation Coefficient ([11],[10]).

As a result of such workflow, for each group up-regulated and low-regulated genes are highlighted and analysed. Section IV will show comparisons and clustering results with such data.

IV. Results

Since our purpose is to investigate how the different strains reacted to the treatments, we first took into account the three pairs for comparison that are the three Sham operated groups against their treated ones: SDP-SDS, ZLT-ZLS, ZDFT-ZDFS (that is, the blu arrows of Fig. 1). The set of differentially expressed (DE) genes was filtered out for each such pair of groups. This resulted already in a quite strong reduction of the research space, because for all pairs the outcome was around few hundreds of genes out of the over 60,000 initial ones (see Figures 2,3,4).

Then, differentially expressed genes resulting from such comparisons were compared between themselves showing a relevant intersection for the set of evolved genes in different strains.

Figure 2 shows the Venn Diagram summarizing results for the pairs ZLT-ZLS against SDP-SDS. There are 476 differentially expressed genes for ZL in the two conditions (Sham and treated), and 464 for SD. As we can see, the intersection between these two sets is very significant. Similarly, Figure 3 shows the set of DE genes in the two conditions for the strains SD and ZDF, and again the intersection is relevant. Finally, Figure 4 shows the sizes and intersection for the DE genes of the strains ZL and ZDF, with the same result of large intersection.

We recall that in all cases the set of differentially expressed genes was extracted using two different techniques (Empirical Bayes-moderated F-statistics with the limma Bioconductor package, and 1 way Anova SNK post hoc test with the GeneSpring Agilent software), and that the results were in both cases this high amount of intersections. We believe that this is a good validation of our model, and an interesting practical result as it allows to bring down the focus from 64,000 genes to a few hundred that, on their turn, can be grouped into very few functional clusters.

To see whether these differences are due to the combination strain/treatments or only to strains we performed also the comparisons: SDS-ZLS, SDS-ZDFS, and ZLS-ZDFS that cross differentially expressed genes data for different strains. Results are shown in Figures 5, 6, and 7. Also in this case, results show a significant intersection, and the combination

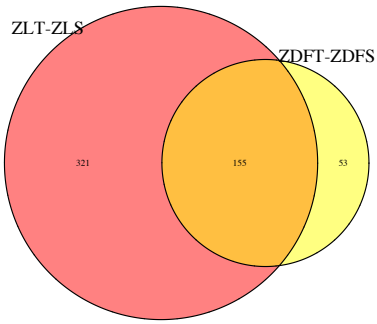

Fig. 4. The right circle represents the set of genes differentially expressed in the two condition (Sham and treatment) for the ZDF Strain, while the left circle is that of the ZL strain in its two conditions.

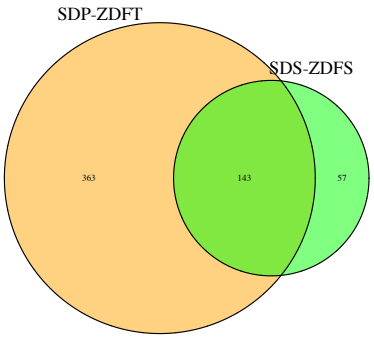

Fig. 5. The left circle represents the set of genes differentially expressed in the treated groups of the strains SD and ZDF, while the right circle is that of the same strains in the Sham operated case.

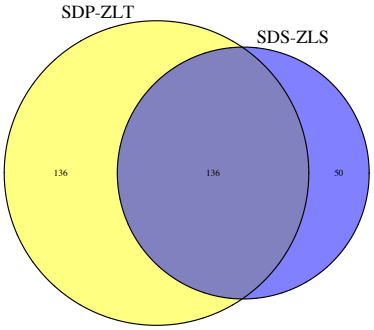

Fig. 6. The left circle represents the set of genes differentially expressed in the treated groups of the strains SD and ZL, while the right circle is that of the same strains in the Sham operated case.

of these and the previous return can drive a more accurate downstream analysis that will follow this work with a focus on the actual set of genes whose differential expression possibly results a significant preclinical result.

Indeed, as a final step, we performed a Gene Ontology analysis, a Functional Clustering analysis, and Pathway analysis on gene lists using, again, different tools and algorithms. We mainly used results from DAVID tools ([12],[17]), and in

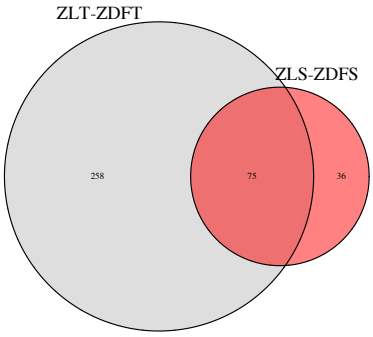

Fig. 7. The left circle represents the set of genes differentially expressed in the treated groups of the strains ZL and ZDF, while the right circle is that of the Sham operated groups of the same strains.

Annotation Cluster 1	Enrichment Score: 12.93			Count	P_Value	Benjamini
GOTERM_BP_FAT	response to wounding	RT		42	4.0E-27	6.7E-24
GOTERM_BP_FAT	inflammatory response	RT		31	2.4E-24	2.0E-21
GOTERM_BP_FAT	defense response	RT		36	1.2E-21	6.5E-19
Annotation Cluster 1	**Enrichment Score: 12.93**			**Count**	**P_Value**	**Benjamini**
GOTERM_BP_FAT	defense response	RT		73	3.4E-38	9.2E-35
GOTERM_BP_FAT	inflammatory response	RT		56	2.9E-37	4.0E-34
GOTERM_BP_FAT	response to wounding	RT		71	1.6E-34	1.5E-31

Fig. 8.

particular from *Functional Annotation Clustering* tool that are shown in Figure 8.

For example, the Functional Annotation Clustering on ZLT-ZLS and on ZDFT-ZDFS differentially expressed genes show the major Enrichment Score for three biological processes: response to wounding, inflammatory response, defense response (Figure 8). These results are not surprising and match the literature concerning the transcriptome gene expression observed in stroke injury. We consider this a validation of our model and of our workflow and results concerning the microarray data analysis. Nevertheless, both the gene expression data itself, and the whole model involving the types of data other than microarray that we collectedt, will be object of future investigations.

V. Conclusions

A first statistical analysis of the data showed different results for the various models, which appear to respond differently to similar treatments. In particular, the analysis of gene expression data showed how the model chosen for conditions that in human being are associated with higher risk for stroke, actually respond differently from other models when subjected to similar treatments.
The use of new kinds of animal models can lead to new results in studies related to cerebral ischemia. Further and more detailed analysis of the database will indicate whether this approach allows a better match between the animal model and human pathological condition.

Acknowledgment

This work was funded by the Project entitled *KEBIC: Creazione di un knowledge base per la biosimulazione e*

l'aumento di predittività e traslazionalità di dati proclitici nel campo dell'ischemia cerebrale of the Regione Toscana, within the call POR CReO FESR 2007-2013 Linea d'intervento 1.5.-1.6, Bando Unico R&S anno 2008.
This work was partially funded by the Project entitled *Metodologie computazionali per la medicina personalizzata* of the University of Pisa, within the call PRA 2015.

REFERENCES

[1] Nichols M1 and Townsend N2 and Scarborough P3 and Rayner M4, *European Cardiovascular Disease Statistics*, 2012 edition. British Heart Foundation Health Promotion Research Group, Department of Public Health, University of Oxford, and Jose Leal and Ramon Luengo-Fernandez, Health Economics Research Centre, Department of Public Health, University of Oxford.

[2] Carden DL1, Granger DN. *Pathophysiology of ischaemia-reperfusion injury* J Pathol. 2000 Feb;190(3):255-66.

[3] Huber W and Carey VJ and Gentleman R and Anders S and Carlson M and Carvalho BS and Bravo HC and Davis S and Gatto L and Girke T and Gottardo R and Hahne F and Hansen KD and Irizarry RA and Lawrence M and Love MI and MacDonald J and Obenchain V and Ole? AK and Pags H and Reyes A and Shannon P and Smyth GK and Tenenbaum D and Waldron L and Morgan M., *Orchestrating high-throughput genomic analysis with Bioconductor.*, 3rd ed. Nat Methods. 2015 Feb;12(2):115-21.

[4] Ritchie ME and Phipson B and Wu D and Hu Y and Law CW and Shi W and Smyth GK, *limma powers differential expression analyses for RNA-sequencing and microarray studies.*, 3rd ed. Nucleic Acids Res. 2015 Apr 20. Epub 2015 Jan 20.

[5] Anders S and McCarthy DJ and Chen Y and Okoniewski M and Smyth GK and Huber W and Robinson MD, *Count-based differential expression analysis of RNA sequencing data using R and Bioconductor.*, 3rd ed. Nat Protoc. 2013 Sep;8(9):1765-86.

[6] McCarthy DJ and Chen Y and Smyth GK, *Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.*, 3rd ed. Nucleic Acids Res. 2012 May;40(10):4288-97.

[7] Bryant PA and Smyth GK and Robins-Browne R and Curtis N., *Technical variability is greater than biological variability in a microarray experiment but both are outweighed by changes induced by stimulation.*, 3rd ed. PLoS One. 2011;6(5):e19556.

[8] Robinson MD and McCarthy DJ and Smyth GK, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.*, 3rd ed. Bioinformatics. 2010 Jan 1;26(1):139-40.

[9] Anders S and Huber W, *Differential expression analysis for sequence count data*, 3rd ed. Genome Biol. 2010;11(10):R106.

[10] Geraci F and Leoncini M and Montangero M and Pellegrini M and Renda ME., *K-Boost: a scalable algorithm for high-quality clustering of microarray gene expression data.*, 3rd ed. J Comput Biol. 2009 Jun;16(6):859-73.

[11] Geraci F1 and Pellegrini M and Renda ME, *AMIC@: All MIcroarray Clusterings @ once.*, 3rd ed. Nucleic Acids Res. 2008 Jul 1.

[12] Olson NE, *The microarray data analysis process: from raw data to biological significance.*, 3rd ed. NeuroRx. 2006 Jul;3(3):373-83.

[13] Smyth GK, *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.*, 3rd ed. Stat Appl Genet Mol Biol. 2004;3:Article3.

[14] Kogure T and Kogure K, Department of Neurosurgery, Tokyo Jikei University School of Medicine, Japan.*Molecular and biochemical events within the brain subjected to cerebral ischemia (targets for therapeutical intervention).*, 3rd ed. Clinical Neuroscience (New York, N.Y.) 1997, 4(4):179-183

[15] Markus HS,Clinical Neuroscience, St George's Hospital Medical School, London, UK. *Cerebral perfusion and stroke.*, 3rd ed. Journal of Neurology, Neurosurgery, and Psychiatry 2004, 75(3):353-361

[16] Koistinaho J and Hkfelt T, A.I. Virtanen Institute, University of Kuopio, Finland. *Altered gene expression in brain ischemia.*, 3rd ed. Neuroreport 1997, 8(2):i-viii

[17] Da Wei Huang and Brad T Sherman and Qina Tan and Jack R Collins and W Gregory Alvord and Jean Roayaei and Robert Stephens and Michael W Baseler and H Clifford Lane and Richard A Lempicki, *The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists*, 3rd ed. Genome Biology 2007, 8:R183