On the Evaluation of the Most Accurate Pediatric Medulloblastoma Animal Model

Behrouz Shamsaei* Cuilan Gao†

Abstract

Animal models of human disease are commonly utilized to gain preclinical insight into the potential efficacy and action mode of novel drugs. The development and selection of an animal model that accurately mimics the human disease profoundly reduces the research timeline and resources needed to make meaningful advances in the treatment and prevention of the human disease under study. Here, we propose a statistical procedure to select the animal model that most accurately mimics the human disease in terms of genome-wide gene expression. Our procedure is designed for studies that have gene expression profiles for a cohort of human disease tissue specimens from different subjects and gene expression profiles for cohorts of disease tissue specimens for each of several animal models. First, we define and compute a metric of similarity between each human gene expression profile and animal gene expression profile which result in multiple groups of similarities. Then a random block ANOVA model is used to compare the group means of similarities between different animal models. Finally post-hot multiple comparison is applied to seek the "best" animal model of the human disease. The advantages of the proposed method are observed in simulation studies and a real example of pediatric Medulloblastoma.

Key Words:
ANOVA, gene expression data, human disease, animal model, Medulloblastoma

1. Introduction

Animal models play a pivotal role in translation biomedical research. The scientific value of an animal model depends on how accurately it mimics the human disease. In principle, microarrays and gene-sequencing technology massive data are commonly used to compare gene expression among biological conditions within the same species. For instance, the experiments may compare the transcriptomes of tumors and host normal tissue or between tumors arising in the same tissue in order to detect the differentially expressed genes. However, translating this wealth of data into other experimental systems has proven difficult because of limitations in comparing transcriptome data generated from different species. Some cross-species gene expression analysis are done in [Ji et al. (2004), Kristiansson et al. (2013) and Johnson et al.(2010)], however statistical methods for cross-species gene expression analysis for this purpose are lacking.

To fill the aforementioned gap, Agreement of Differential Expression Analysis (AGDEX) package is developed by Pounds et al. (2011). AGDEX is a method that detects geno-wide transcriptomic similarities in gene expression between tissues from human and animal by comparing the gene expression of shared ortholog genes. By using cosine and Difference of Proportions similarity metrics, AGDEX evaluates the level of similarity between human tissue and animal models. Yet AGDEX can not detect which animal model is the most accurate one that can mimic the human disease among a set of animal models, which is a more crucial question. Take the pediatric medulloblastoma study [Kawauchi et al. (2012)] for example, there are 4 subgroups of medulloblastoma and they differ in histo-pathology, gene expression profile, and clinical behavior from other forms. Thus cardinal features of mouse medulloblastomas that can closely mimic those

*PhD. Candidate, Department of Computational Engineering, University of Tennessee at Chattanooga, 701 E. M. L. King Blvd, Chattanooga, TN 37403

†Assistant Professor of Statistics, Department of Mathematics, University of Tennessee at Chattanooga, Dept 6956, 615 McCallie Ave, Chattanooga, TN 37403

subgroups of human medulloblastomas need to be identified. For a specific subgroup of medulloblastoma, which animal model is the most accurate one is another important question to ask as a following up question of AGDEX. Here we propose a statistical procedure to identify the most accurate model among a set of candidate animal models via calculating the pairwise similarities between human and animal tissues and then compare the seminaries among different models.

2. Methodology

In practice number of human gene expression samples for a particular type of disease is limited and we call these set of samples, a human model. Defining that, each human or animal model is a span of several gene expression samples of a particular disease. So human model can be illustrated as

$$\overbrace{\left\{\overbrace{\begin{pmatrix} gene-1 \\ gene-2 \\ gene-3 \\ \vdots \\ gene-n \end{pmatrix}}^{sample1} \overbrace{\begin{pmatrix} gene-1 \\ gene-2 \\ gene-3 \\ \vdots \\ gene-n \end{pmatrix}}^{sample2} \overbrace{\begin{pmatrix} gene-1 \\ gene-2 \\ gene-3 \\ \vdots \\ gene-n \end{pmatrix}}^{sample3} \dots \overbrace{\begin{pmatrix} gene-1 \\ gene-2 \\ gene-3 \\ \vdots \\ gene-n \end{pmatrix}}^{sample-h}\right\}}^{\text{Human model}} \tag{1}$$

Relation (1) shows the human model with h samples where each sample consists of n number of genes. We can provide several animal models of different types of diseases, each of which consisting of several animal samples. Relation (2) shows the i^{th} animal model with m samples, each including n genes.

$$\overbrace{\left\{\overbrace{\begin{pmatrix} gene-1 \\ gene-2 \\ gene-3 \\ \vdots \\ gene-n \end{pmatrix}}^{sample1} \overbrace{\begin{pmatrix} gene-1 \\ gene-2 \\ gene-3 \\ \vdots \\ gene-n \end{pmatrix}}^{sample2} \overbrace{\begin{pmatrix} gene-1 \\ gene-2 \\ gene-3 \\ \vdots \\ gene-n \end{pmatrix}}^{sample3} \dots \overbrace{\begin{pmatrix} gene-1 \\ gene-2 \\ gene-3 \\ \vdots \\ gene-n \end{pmatrix}}^{sample-m}\right\}}^{\text{Animal model}-i} \tag{2}$$

In this study we have assumed we have several number of animal models and only one human expression profile and the purpose of this research is to find the most similar animal model to the human one.

2.1 ANOVA Models

Analysis of variance(ANOVA) is used to find the effects of categorical independent variables (factors) on associated dependent variables. The very first step is to introduce ANOVA models to reach our goal. For this study two ANOVA models are proposed. In the first ANOVA model, effect of similarity of animal models to the human model is regarded as a fixed factor and in the second ANOVA model effect of random block human samples is regarded as well.

$$\left\{\begin{array}{l} First \text{ ANOVA } model: Y_{im} = \eta + \alpha_m + \varepsilon_{im} \\ Second \text{ ANOVA } model: Y_{im} = \eta + \alpha_m + h_i + \varepsilon_{im} \end{array}\right\} \left\{\begin{array}{l} 1 < m < number\,of\,animal\,models \\ 1 < i < number\,of\,human\,samples \end{array}\right. \tag{3}$$

where η is the grand mean, α_m is the effect of m^{th} factor which is the similarity of m^{th} animal model to the human model, h_i is the effect of i^{th} human random block samples and ε_{ij} is the associated error. h_i can be regarded as a block factor since the human data can be obtained from different geographical population or different races and genders. To define the ANOVA models we have to define similarity metrics between pairwise animal models and the human model, so the proceeding sections are devoted to define the Y_{im}, i^{th} treatment in the m^{th} factor. The factors are correlation vectors between human and animal models. Null hypothesis or H_0 is if there is no significant effect of animal models, i.e. $\alpha_1 = \alpha_2 = ... = \alpha_m$, where α_i is the effect of i^{th} animal model and alternative hypothesis orH_α is if there is a significant difference between animal factors.

2.2 Metrics of similarity

To define the ANOVA models, we have to determine a metric showing how similar the animal models to the human is. As discussed above each of these models is a matrix, so in the following section three different metrics are presented showing the similarity between the columns of the matrices or the projection of columns of human model on the span of animal model matrices.

2.2.1 Semi-Correlation(S-Cor)

One way to show the correlation between two matrices is to find the correlation between the columns of the two matrices. Experience shows that this method may result in non-normally distributed data. So to remedy this deficiency a new metric is defined baesd on correlation coefficient between i^{th} human sample and the samples of m^{th} animal model. Defining this, Y_{im} in relation (3) can be defined as

$$Y_{im} = \sum \left(h_i - \bar{h}_i\right) \times \left(a_{m,j} - \bar{a}_{m,j}\right)^T \tag{4}$$

Where in equation (4) $a_{m,j}$ is the j^{th} sample of m^{th} animal model, h_i is the i^{th} human sample, $\bar{*}$ refers to mean of variable $*$ and $*^T$ denotes the transpose of $*$. Likewise we can define other metrics.

2.2.2 Cosine(Cos)

Second metric, like the first metric, defines a similarity between human and and animal samples, which are columns of human model and each animal model. This metric is also used by Pounds et al. (2011) and characterizes the $Cosine$ between two vectors.

$$Y_{im} = \frac{\sum \left(h_i \times a_{m,j}^T\right)}{\sqrt{\sum \left(h_i\right)^2} \times \sqrt{\sum \left(a_{m,j}\right)^2}} \tag{5}$$

Definition of parameters in equation (5) is like the counterparts in equation (4).

2.2.3 Projection(Pro)

This method is founded on the concept of projection of the vectors on a matrix span. One can claim that if a vector is closer to the span of a matrix, larger norm of projection of the vector to span of matrix is expected. The projection vector is a linear least squares data-fitting solution that can be written as

$$A \times x \approx b \tag{6}$$

where b is the vector, A is the matrix and x is the least square solution, Also $A \times x$ is the projection vector. In our particular application, A is analogous to an animal model and b represents a human sample.

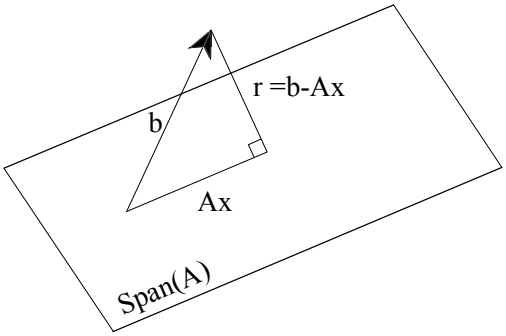


Figure 1: Geometric depiction of projection

In figure 1, r defines the residual, because the vector b, may not reside in span of matrix A. Since the matrix A is not a square matrix, we have to employ least square solutions. To preserve the Euclidean norm of matrix during the solution, we have used the orthogonal transformation to solve the equation (6).

Given an $m \times n$ matrix A, with $n \leq m$, we seek a $m \times m$ orthogonal matrix Q such that

$$A = Q \begin{bmatrix} R \\ 0 \end{bmatrix} \tag{7}$$

where R is $n \times n$ and upper triangular matrix. Such a QR factorization transforms the linear least squares problem $Ax \approx b$ into a triangular least squares problem having the same solution, because

$$\|b - Ax\|_2 = \left\| b - Q \begin{bmatrix} R \\ 0 \end{bmatrix} x \right\|_2 = \left\| Q^T b - \begin{bmatrix} R \\ 0 \end{bmatrix} x \right\|_2 \tag{8}$$

Where $\|*\|_2$ represents the Euclidean norm of $*$. The solution to this problem is

$$x = R^{-1} Q^T b \tag{9}$$

And finally the independent variable Y_{im} is the norm of projection of i^{th} human sample on the m^{th} animal model span. which is

$$Y_{im} = \|A \times x\|_2 \tag{10}$$

2.3 Post hoc analysis

Defining the ANOVA models in relation(3), we first check if the group means of similarities between different animal models and the human model are different. Typically bigger mean values show more similarity of the animal model to the human counterpart.

Another way of finding the most similar model is using results of ANOVA table and check the ANOVA assumptions. And finally use of multiple comparisons procedure test(Tukey's test or Hsu's Best) to identify the best animal model.

3. Examples and Results

So far two ANOVA models and three metrics of similarities to build the ANOVA models are introduced. To check the capabilities of the proposed schemes two different examples are illustrated. In the first example, a set of manufactured data is generated and described process in the previous section is applied on. For the second example, ANOVA procedure is implemented on pediatric brain tumor data.

3.1 Simulated data

To mimic the real data, some normally distributed data is generated for one human model, with h samples and n number of genes, and for k animal models each with m samples and n number of genes. The first animal model would be the summation of a function of human model and an error term to simulate the most similar animal model to human one. ANOVA models should be applied to check the null hypothesis. We should check the ANOVA assumptions as well.

3.1.1 Generated human and animal models

As discussed before, a set of random numbers is generated for the both human and animal data. For the human model h number of samples each with 1000 gene data is generated with gene expression values between 0 to 1. K number of samples with m number of samples with the same number of genes are also similarly generated. First animal model is designed as

$$A_1 = \frac{\rho}{\sqrt{1-\rho^2}} \times H + E \tag{11}$$

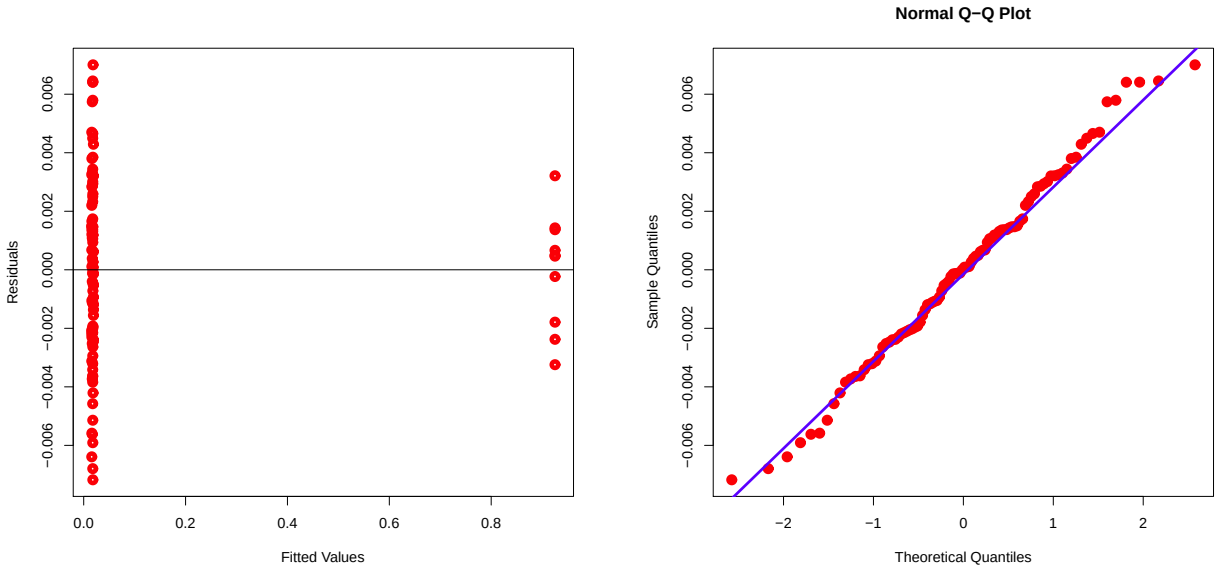
Where A_1 is the first animal model, H is the human model, ρ is a scaling factor between 0 to 1, $(0 < \rho < 1)$ and E is a user defined error term, here defined random numbers between 0 to 1. Arbitrary number of human and animal samples and animal models can be generated with same number of genes, but for the illustration purpose, number of examples is restricted to three and specified in table 1.

Table 1: Parameter settings

Examples	ρ	Human samples	Animal samples	Animal models	number of genes
$Example1$	0.0, 0.1, 0.2, ..., 0.9	5	5	3	1000
$Example2$	0.6	5, 10, 15, 20	5	3	1000
$Example3$	0.6	20	5, 10, 15, 20	3	1000

3.1.2 ANOVA Assumption check list

For all of these examples, minimum p-values for testing the assumptions of ANOVA, Shapiro test and Bartlet test are 0.1, resulting in satisfaction of ANOVA assumptions, i.e. normality of data and homogeneity of variance. Figure 2 on the next page show the fitted value and Q-Q plot of results for the first example of second method.



(a) Plotting fitted values of first sample of first setting

(b) Q-Q plot of first sample of first setting

Figure 2: Graphical illustration of ANOVA assumptions

3.1.3 Results

Accumulative means of the proposed examples are tabulated in tables 2, 3 on the next page and 4 on the following page. These tables are generated for the first proposed ANOVA model that only includes the animal effects. These tables show that means of similarity in the first model are significantly higher than the means of other models.

Table 2: Accumulative mean in groups in Setting 1

Setting			Model1			Model2			Model3		
ρ	h	m	$S-Cor$	Cos	Pro	$S-Cor$	Cos	Pro	$S-Cor$	Cos	Pro
0	5	5	197.71	0.19	8.21	105.03	0.11	7.05	107.93	0.11	6.75
0.1	5	5	228.32	0.23	31.91	105.03	0.11	7.05	107.93	0.11	6.75
0.2	5	5	259.87	0.25	32.14	105.03	0.11	7.05	107.93	0.11	6.75
0.3	5	5	293.48	0.27	32.35	105.03	0.11	7.05	107.93	0.11	6.75
0.4	5	5	330.61	0.287	32.52	105.03	0.11	7.05	107.93	0.11	6.75
0.5	5	5	373.52	0.30	32.66	105.03	0.11	7.05	107.93	0.11	6.75
0.6	5	5	426.10	0.312	32.79	105.03	0.11	7.05	107.93	0.11	6.75
0.7	5	5	496.19	0.321	32.91	105.03	0.11	7.05	107.93	0.11	6.75
0.8	5	5	603.73	0.325	33.03	105.03	0.11	7.05	107.93	0.11	6.75
0.9	5	5	826.44	0.322	33.14	105.03	0.11	7.05	107.93	0.11	6.75

Table 3: Accumulative mean in groups in Setting 2

Setting			Model1			Model2			Model3		
ρ	h	m	$S-Cor$	Cos	Pro	$S-Cor$	Cos	Pro	$S-Cor$	Cos	Pro
0.6	5	5	426.10	0.312	32.79	105.03	0.11	7.05	107.93	0.11	6.75
0.6	10	5	267.88	0.196	18.82	109.49	0.109	6.89	111.97	0.109	7.25
0.6	15	5	209.76	0.153	13.95	106.30	0.105	6.98	94.33	0.093	6.34
0.6	20	5	182.66	0.133	11.45	94.69	0.093	6.35	95.11	0.097	6.52

Table 4: Accumulative mean in groups in Setting 3

Setting			Model1			Model2			Model3		
ρ	h	m	$S-Cor$	Cos	Pro	$S-Cor$	Cos	Pro	$S-Cor$	Cos	Pro
0.6	20	5	182.66	0.133	11.45	94.69	0.093	6.35	95.11	0.097	6.52
0.6	20	10	181.87	0.136	19.09	96.65	0.097	8.18	102.07	.103	8.33
0.6	20	15	179.99	0.136	26.32	100.78	0.101	9.18	101.24	.103	9.13
0.6	20	20	180.58	0.138	32.92	100.21	0.102	9.76	105.82	0.105	10.33

ANOVA table shows the rejection of null hypothesis. Table 5 shows the results of the first ANOVA model for the last sample in example 3 for the Cosine metric.

Table 5: ANOVA table results for the first sample of third setting with cosine metric

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
$model$	2	0.319	0.15962	11.74	8.92e-06
$Residuals$	1197	16.273	0.01359		

And finally results of Turkey test for the same sample in table 6 shows that the first factor $model1$ is different from the other factors $model2$ and $model3$, from which we can conclude the first animal model is the most similar model to human one.

Table 6: Results of Turkey test for the first sample of third setting with cosine metric

	diff	lwr	upr	p adj
$model2-model1$	-0.035644321	-0.05499123	-0.01629741	0.0000494
$model3-model1$	-0.033450432	-0.05279734	-0.01410352	0.0001562
$model3-model2$	0.002193888	-0.01715302	0.02154080	0.9617196

Analysis show the random block human samples are not a significant factor and it suffices the use of only first ANOVA model.

3.2 Real data

These data is gathered from pediatric brain medulloblastoma tumor and can be found in NCBI by access number GSE33199 and GSE33200. medulloblastoma Human data consists of 106 samples with 54675

probe sets(genes). Mouse data consists of 4 models or subgroups that they differ in histo-pathology, 19 total samples and each sample contains 45102 probe sets. Animal data are categorized in four subgroups of normal: 5 samples, stem: 5 samples, prog: 5 and ptch 4 samples, 19 samples total.

A mapping procedure is induced to find the counter-probe set of human data in the mouse data. From the proposed similarity metrics the first two metrics finds the correct cancer type (ptch), the reason that the "Projection" metric gives the wrong result is because of small number of animal samples and high level of the singularity in the animal matrix spans. The results of pair-wise t-test of the Semi-Correlation and Cosine metrics are tabulated in tables 7 and 8, respectively.

Table 7: Pair-wise t-test of real data with semi-correlation metric

	$norm$	$prog$	$ptch$
$prog$	0.012	-	-
$ptch$	1.2e-12	1.9e-06	-
$stem$	0.035	0.686	2.7e-07

Table 8: Pair-wise t-test of real data with cosine metric

	$norm$	$prog$	$ptch$
$prog$	0.87122	-	-
$ptch$	7.6e-05	0.00014	-
$stem$	0.73419	0.85915	0.00028

4. Conclusion

One way to find the human disease is to statistically match the genomic data between human and animal gene expression models. Each animal and human model consists of several samples. In this research two ANOVA models with three different metrics of similarities are presented to find the the best animal model that mimics the human disease genomic model. The first ANOVA model only checks the significance of similarity between human and animal models and the second ANOVA model is similar to the first model but considers the human samples as a random block variable as well.

The first similarity metric is based on the correlations between each animal and human samples. Concept of cosine between two vectors is regarded as the second similarity metric and finally the third proposed metric is based on the projection of human sample vectors on animal models span. Some simulation examples are illustrated to show the efficacy of the proposed methods and finally real genomic data is used to check the potentiality of the schemes.

Results show that first two metrics, semi-correlation and cosine of similarities are the better choices to find the similarities, because the third metric highly depends on the level of singularity of the animal models span, which in the real data example are very high. Also outcome of ANOVA models for these examples and set of data reveals the insignificance of human samples as random block variables that suffices the use of first ANOVA model. The proposed method is a framework for identifying the most accurate animal model of human disease which can be applied to other similar studies.

REFERENCES

Ji, W., Zhou, W., Gregg, K., Yu, N., Davis, S. and Davis, S.(2004), "A method for cross-species gene expression analysis with high-density oligonucleotide arrays," *Nucleic Acids Research* , Vol. 32, No. 11, e93.

Kristiansson, E., Österlund, T., Gunnarsson, L., Arne, G., Larsson, D.G.J. and Nerman, O. (2013)," A novel method for cross-species gene expression analysis," *BMC bioinformatics* , 14:70.

Pounds, S., Gao, C. L., Johnson, R. A., Wright, K. D., Poppleton, H., Finkelstein, D., Leary, S. E. S. and Gilbertson, R. J. (2011)," A procedure to statistically evaluate agreement of differential expression for cross-species genomics," *Bioinformatics*, 27(15), pp. 2098-103.

Kawauchi, D., Robinson, G., Uziel, T., Gibson, P., Rehg, J., Gao, C., Finkelstein, D., Qu, C., Pounds, S., Ellison, D. W., Gilbertson, R. J. and Roussel, M. F. (2012)," A mouse model of the most aggressive subgroup of human medulloblastoma," *Cancer Cell*, 21(2), pp. 168-80.

Johnson, R. A., Wright, K. D., Poppleton, H., Mohankumar, K. M., Finkelstein, D., Pounds, S. B., Rand, V., Leary, S. E., White, E., Eden, C., Hogg, T., Northcott, P., Mack, S., Neale, G., Wang, Y. D., Coyle, B., Atkinson, J., DeWire, M., Kranenburg, T. A., Gillespie, Y., Allen, J. C., Merchant, T., Boop, F. A., Sanford, R. A., Gajjar, A., Ellison, D. W., Taylor, M. D., Grundy, R. G. and Gilbertson, R. J. (2010)," Cross-species genomics matches driver mutations and cell compartments to model ependymoma," *Nature* , 466(7306), pp. 632-6.