# Comparability of Inductive Rules of Gene Enrichment Between Humans and Mice

Galen Michael Seilis
*Department of Computer Science*
*University of Northern British Columbia*
Prince George, Canada
seilis@unbc.ca

***Abstract*—Mice are used as model organisms for understanding other vertebrates in thousands of studies, including to better understand humans. Gene enrichment analysis has become a popular approach to analyzing patterns in genes, that has led to publicly-available datasets that are suitable to an association rule learning approach. In this study we found that there are dozens of biological processes where inferences about mice accurately apply to humans, and that thousands of biological processes are relatively less reliable.**

***Index Terms*—association rule learning, fpgrowth, confusion matrix, exploratory data analysis**

## I. Introduction

Genes in eukaryotic organisms such as humans and mice are sections of DNA that regulate and encode biological functions through processes such as transcription, translation, and post-translational modifications. For our purposes here, transcription is the production of RNA from DNA, translation is the production of protein from RNA, and post-translational modifications is a miscellanous category for alterations to a protein after its production. As many biological processes are performed through the activity of proteins, RNA, the flow of information I just described sets for us an expectation that genes are associated with biological functions. These associations are many-to-many as a single gene will participate in many biological processes, and a single biological process is influenced by many genes.

Many genes usually participate in any given biological process, and the role of a protein encoded by a given gene may be very specific and even subtle. For example, the majority of proteins involved in $\beta$-oxidation of fatty acids are *not* catalyzing oxidation reactions but rather dehydrogenation, hydration, and thiolysis.[1] Genes participating in the same biological process may have relations to each other. In the example of $\beta$-oxidation, the gene that encodes 3-hydroxyacyl-CoA dehydrogenase which catalyzes the oxidation step of the processes requires that the enzyme enoyl CoA-hydratase (another protein encoded by a different gene) has catalyzed the hydration any trans-$\Delta^2$ bonds in the enoyl-CoA intermediate.[1]

From the last two paragraphs, I've given precedent to the idea that genes are associated with each other because of their complementary roles in biological processes. While it does not hold exactly true in every case, it is generally true that organisms that are more evolutionarily related (i.e. shared a more recent common ancestor) will be more similar in their genes and therefore biological processes. In medical studies, mice are often used as an early model of humans in understanding human behaviour, toxicology, or drug development.[2, 3, 4] It would be potentially yield useful insights to examine how representative mice are of humans in terms of gene participation in biological processes.

Frequent pattern mining and association rule mining are tasks commonly found in data mining for knowledge discovery in databases. Given a transaction database where each transaction is a biological process and the items are genes involved in that process, associations between genes should be discoverable in the form of either frequent patterns or association rules. Assocation rules are preferable because could they can represent asymmetric relations between genes. Biological databases are available for humans and mice, and mining of those datasets, can be used to assess if mice are representative of humans in terms of association rules.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a biochemical database that is rich in data about genes, chemical reactions, pathways, and biological processes.[5] KEGG's application programming interface (API) makes gene enrichment data available, but a growing trend of gene enrichment analysis has led researchers to use curated database like Enrichr for easier access to gene enrichment datasets.[6] A gene enrichment dataset is easily reconceived of as a transaction database as described above because it simply lists which genes are involved in which biological processes. The Enrichr project has collated such dataset from KEGG for both humans and mice. This study will look at what agreement in association rules can be found between humans and mice.

## II. Methodology

This section will explain the process of how the data was obtained and analyzed in this study.

The first step in this analysis was obtaining the required datafiles, which was a simple matter of downloading the tab-delimited text files from the Enrichr database interface. This dataset is formatted such that each biological process can be represented as a transaction, with its itemsets being the collection of genes that participate in that biological process.

There are numerous frequent pattern mining algorithms available, of which all are probably capable of processing the 308 human transactions and 303 mouse transactions in the KEGG Enrichr databases. In this study we used the FP-growth algorithm implemented from the MLxtend because of its smooth integration with other common Python libraries such as Pandas.[7, 8] The Pandas dataframe data structure is especially versatile and performant in applying mathematical and SQL-like operations.[9] A minimum support threshold of 0.18 was used because it was a balance between the tradeoff of runtime of the implementation with getting numerous frequent itemsets out of a small dataset. Once these frequent patterns were obtained, association rules were calculated from them using a minimum confidence threshold of 1.0 because we were interested in only the most strongly-inductive rules.

With the frequent patterns and association rules mined from the databases, a comparative analysis was performed between the association rules of humans and mice using a confusion matrix evaluation method.

Beyond the interspecies comparative analysis of association rules found within gene enrichment, structure within the data were also explored using principal component analysis and visual examination of a graph representation of the genes.

## III. Results and Discussion

This section will exposit the results of the analyses performed, and will discuss potential implications of those results.

### A. Comparative Analysis of Association Rules

The set of association rules extracted from the human gene enrichment data was not identical to that of the mouse, however there was a large enough overlap to consider a bivariate statistical comparison. Figure 1 (top panel) shows that for the association rules that were extracted from both species have strongly-correlated supports. While it is true that *correlation does not imply causation*, Reichenbach's principle suggests that there should be some underlying reasons for a stable correlation to exist.[10] One hypothesis is that the genes participating in the biological processes that these rules were extracted from are homologs under strong selective pressure that prevents them from diverging, yielding similar supports in the same biological processes. This hypothesis is untested by the current study, but might be confirmed by modern phylogenetic estimation techniques such as Bayesian Evolutionary Analysis by Sampling Trees (BEAST).[11]

Association rules are analogous to an implication statement in that $A \to B$ for itemsets $A$ and $B$ describe a dependence of $B$ on $A$. However, an even more apt interpretation of such association rules is the notion of conditional probability from an itemset event space where $A \to B$ as above would be interpreted as $P(E_B|E_A)$.[12, 13] The itemset $B$ is often called the *consequent*, and the itemset $A$ the *antecendent*.[13] The consequent and antecedent themselves are frequent itemsets with supports, and their relationship is explored in the bottom two scatterplots of Figure 1. Both scatterplots show a weak but statistically significant correlation between the support of the antecedents and consequents, which is conditioned by having been filtered by the minimum support threshold. This leads to an reverse-upper-triangular region of the plots containing points, but does not have any biological meaning. The correlation is weak because their is high variance in both support values, which may speak to divergent evolution in how these biological processes progress.

With association rules for both species, it was possible to calculate confusion matrix statistics that evaluate to what extent rules about one species tell us about the rules in the other. These statistics were calculated in two different ways because of a numerical issue that was encountered, indicated by the headings *Min Bounded* and *Max Bounded* in Table I. Confusion matrix statistics take in the number of true positives, the number of false positives, the number of false negatives, and the number of true negatives. It is the last entry that led to calculating the statistics two different ways because it was either zero or a very large number depending on how to interpret the what the population under consideration is. Given a transaction database that contains $n$ number of items among its transactions, Equation 1 gives the number of possible association rules $R$ that can be calculated from such a database in principle.[14]

$$R = 3^n - 2^{n+1} + 1 \tag{1}$$

Using Equation 1, we can use this equation to calculate the number of true negatives by subtracting the number of extracted association rules from it which is obtained using the inclusion-exclusion principle. Concretely, the true negatives can be calculated with Equation 2 under the assumption of considering the true negatives to be the set of all conceivable association rules that were not extracted in an analysis. This large number of $\sim 10^{4747}$ false negatives is the condition presumed in the *Max Bounded* column of Table I.

$$\text{True Negatives} = R_{all} - (R_{\text{mice}} + R_{\text{human}} - R_{\text{mice} \cap \text{human}}) \approx 10^{4747} \tag{2}$$

An alternative interpretation is that the confusion matrix is applied only to the association patterns that were extracted from the database, which leads to the number of true negatives being zero. This interpretation is the presumed condition in the *Min Bounded* column of Table I.

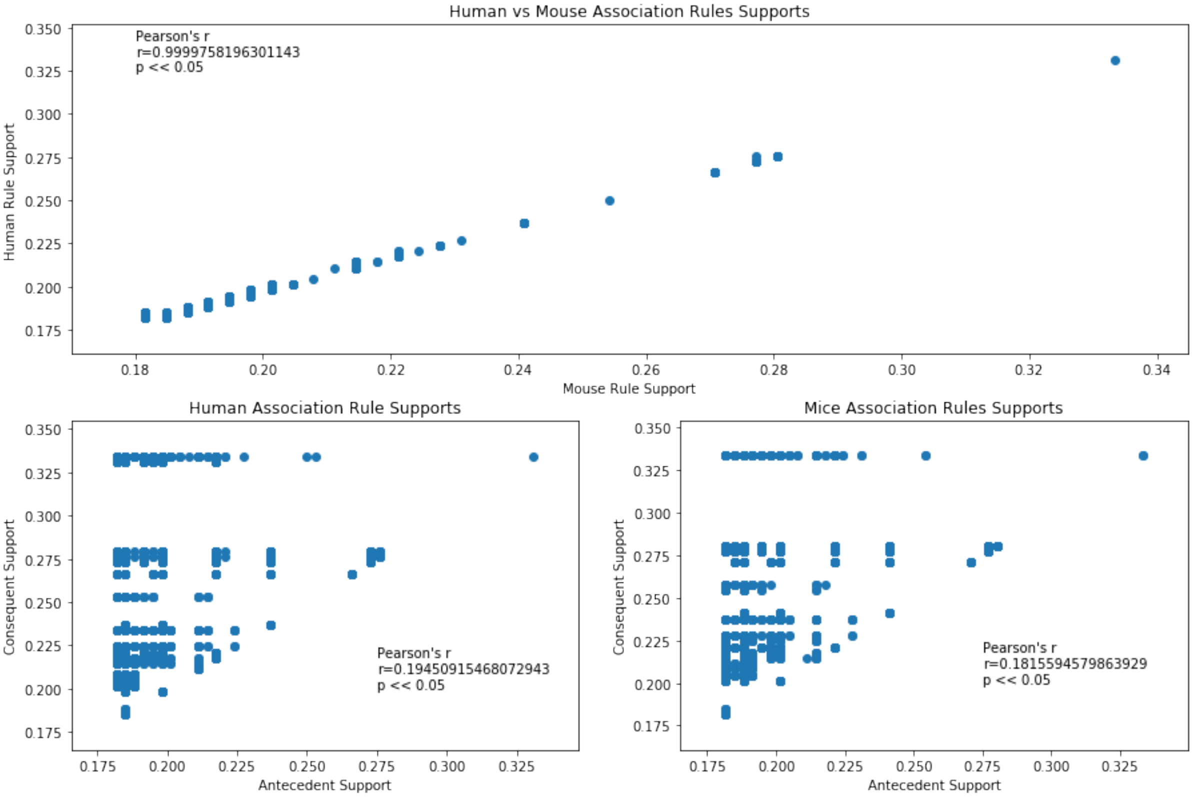


Fig. 1. Comparison of association rule support between humans and mice. (top) Scatterplot comparing the support of association rules of mice to that of humans that were discovered in both species. There is a strong and statistically significant correlation between these variables. (bottom left) Scatterplot showing the relationship between the support of the antecedents or each human gene enrichment association rule to that of their consequent. There is a weak but statistically significant correlation between these variables. (bottom right) Scatterplot showing the relationship between the support of the antecedents or each mouse gene enrichment association rule to that of their consequent. There is a weak but statistically significant correlation between these variables.

This is a finite dataset that should intuitively lead to finite conclusions, however some of the confusion matrix statistics calculated in Table I defy this intuition in both the *Min Bounded* and *Max Bounded* columns by yielding inifinite results. We interpret these results to mean *very large but finite* (i.e. $\sim 10^{4747}$) because of limited floating point precision, rather than the statistics being actually representing the measure or cardinality of an infinite set.

Since the general context of this analysis is to learn how reliable inferences are about humans given what we learn about mice, the ”ground truth” of the confusion matrix represent the human results and the ”predicted condition” represents the mouse results.

TABLE I
CONFUSION MATRIX EVALUATION OF RULE CORRESPONDANCE BETWEEN HUMANS AND MICE

| Statistic | Min Bounded | Max Bounded |
|---|---|---|
| Total Population | 204706.0 | inf |
| Positive | 200205.0 | 200205.0 |
| Negative | 4501.0 | inf |
| Predicted Condition Positive | 204538.0 | 204538.0 |
| Predicted Condition Negative | 168.0 | inf |
| True Positive | 200037.0 | 200037.0 |
| False Positive | 4501.0 | 4501.0 |
| False Negative | 168.0 | 168.0 |
| True Negative | 0.0 | inf |
| Prevalence | 0.978 | 0.0 |
| Accuracy | 0.977 | 1.0 |
| Precision | 0.978 | 0.978 |
| False Discovery Rate | 0.022 | 0.022 |
| False Omission Rate | 1.0 | 0.0 |
| Negative Predictive Value | 0.0 | 1.0 |
| Recall | 0.999 | 0.999 |
| False Negative Rate | 0.001 | 0.001 |
| False Positive Rate | 1.0 | 0.0 |
| Specificity | 0.0 | 1.0 |
| Positive Likelihood Ratio | 0.999 | 3.990634567558178e+304 |
| Negative Likelihood Ratio | inf | 0.001 |
| Diagnotistic Odds Ratio | 0.0 | inf |
| F1 Score | 0.988 | 0.988 |
| Prevalence Threshold | 0.5 | 0.0 |
| Threat Score | 0.977 | 0.977 |
| Balanced Accuracy | 0.5 | 1.0 |
| Matthews Correlation Coefficient | -0.004 | nan |
| Fowlkes Mallows Index | 0.989 | 0.989 |
| Bookmaker Informedness | -0.001 | 0.999 |
| Markedness | -0.022 | 0.978 |
| Fishers Exact Test P | nan | nan |

The first statistic is the total population, which was 204706 in the minimum bounded case and arbitrarily large in the maximum bounded case. Among these, 200205 of them were positive in both cases. The number of these that are negative are 4501 in the minimum bounded case, and arbitrarily large in the maximum bounded case. This simply reflects that there were 4501 negatives (i.e. human and mice association rules didn't match) before considering all conceivable association rules. The number of predicted condition positives (i.e. the number of association rules for mice), was 204538 in both cases. The number of predicted condition negatives was 168 in the minimum bounded case, and arbitrarily large in the maximum bounded case. This means that number of associations rules that were not found for mice was only 168, which is substantially smaller than the number that were found for mice.

The prevalence, $\frac{\sum \text{Condition Positive}}{\sum \text{Total Population}}$, represents the probability that a randomly-selected association rule would be found in the set of the human association rules. In the minimum bounded case, it was 97.8%, because the large majority of association rules were found in both humans and mice among those observed. In the maximum bounded case, the number of observed is less than a percent of those that are conceivable.

The accuracy, $\frac{\sum \text{True Positive} + \sum \text{True Negative}}{\sum \text{Total Population}}$, represents the frequency when either both humans and mice shared association rules or both didn't have a given association rule. This measure is important because it quantifies agreement between the human results and mouse results. In the minimum bounded case, the accuracy was 97.7%, suggested a high level of agreement. The maximum bounded case was $\sim$ 1.0, which is the highest value possible for this measure. Unlike some other measures, accuracy between the minimum bounded case and maximum bounded case are in agreement of having high accuracy. This implies that the accuracy is high regardless of which interpretation is made, and therefore it is more robustly the case that there is a large agreement between the association rules of humans and mice.

The precision, $\frac{\sum \text{True Positive}}{\sum \text{Predicted Condition Positive}}$, is the conditional probability that a randomly selected association rule is a true positive given that it was predicted to be positive. In terms of this study, this measures to what extent we can expect a rule to hold for humans given that it held for mice. Because this score doesn't depend on the true negative count, it was the same in both the minimum and maximum bounded cases. In both cases, the precision was 97.8%, suggesting that a rule holding for mice is probably true descriptive of humans whether it with respect to within-sample counts or within the conceivable space of rules for the given collection of genes.

The false discovery rate, $\frac{\sum \text{False Positive}}{\sum \text{Predicted Condition Positive}}$, is the conditional probability of a randomly selected rule being false given that it was predicted to be true. For this study, the false discovery rate was 2.2% in both bounding cases, suggesting there is a non-trivial but low probability a rule being predicted as true for mice will turn out to be false in humans.

The false omission rate in this study, $\frac{\sum \text{False Negative}}{\sum \text{Predicted Condition Negative}}$, is the conditional probability that a rule which did not hold in mice will also not hold in humans. This measure is in exact disagreement between the minimum bounded case and the maximum bounded case, with the prior indicating *certainly yes* and the latter indicating *certainly no*. Since the probability is decided entirely by the interpretation, it only tells us that the interpretation itself is important.

The negative predictive, $\frac{\sum \text{True Negative}}{\sum \text{Predicted Condition Negative}}$, is the conditional probability that a rule will not hold in humans given that it held in mice. Echoing reverse results of the comments made above for the false omission rate, the minimum bounded case gives a 0 probability, and the maximum bounded case gives a 1 probability. This again suggests that depending on whether you consider only the sample, or the hypothetical space of rules that could have been inducted in principle, has a huge impact on interpreting these confusion matrix statistics.

The recall, $\frac{\sum \text{True Positive}}{\sum \text{Condition Positive}}$, represents the probability of that a rule that worked for mice also works for humans given that the rule does work for humans. This is a measure of the sensitivity of the mouse model in predicting inductive rules for humans. In both the minimum bound and maximum bounded case, this probability was 99.9%, suggesting that whether a rule holds for humans is sensitive to what rules hold for mice.

The false negative rate, $\frac{\sum \text{True Positive}}{\sum \text{Condition Positive}}$, represents the probability that a rule does not hold for mice but does hold for humans. This is more intuitively the *miss rate* of using mice association rules to predict human association rules. With a false negative rate of only 0.1% in both the bounded cases, it is very unlikely that such misses will occur.

The false positive rate, $\frac{\sum \text{False Positive}}{\sum \text{Condition Negative}}$, is the probability that an inductive rule held for mice but didn't hold for humans given that the rule doesn't hold for humans. In the minimum bounded case, this was certainty, and in the maximum bounded case it had a zero probability. Like some of the other measures, the interpretation dominates this measure's value.

The specificity, $\frac{\sum \text{True Negative}}{\sum \text{Condition Negative}}$, represents the conditional probability of a rule not holding for mice and humans given that it doesn't hold for humans. Intuitively, this is a measure how selective using a mouse model is to represent human association rules. Because it is a function of the true negative rate, it is dominated by the interpretation of which space is being used, with zero specificity in the minimum case and one in the maximum case.

The positive likelihood ratio, $\frac{\text{True Positive Rate}}{\text{False Positive Rate}}$, is the comparison of how often mice inductive rules predict human inductive rules compared to when they don't. The minimum bounded case indicates that they are approximately equal, while the maximum bounded case suggests that the true positive rate is much larger.

The negative likelihood ratio, $\frac{\text{False Negative Rate}}{\text{True Negative Rate}}$, is a comparison of how frequently a rule doesn't hold for mice when it holds for humans compared to how frequently a rule doesn't hold for humans or mice. In this study the minimum bounded case suggested that the former was much larger than the latter, while the maximum bounded case suggested that the opposite was true.

The diagnostic odds ratio, $\frac{\text{Positive Likelihood Ratio}}{\text{Negative Likelihood Ratio}}$, combines the previously-discussed likelihood ratio to produce a comparison of the product of the true result rates to the product of the false result rates. This result is interpretation-dependent because both of the ratios it is calculated from are interpretation-dependent, with the minimum bounded case giving a value of zero and the maximum bounded case giving an indefinitely large value.

The F1 score, $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$, is the harmonic mean of the precision and recall. This measure assumes that precision and recall are equally important, but this is not true in general. There is a parametrization that can balance their relative importance given in equation 3.[15]

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}} \tag{3}$$

The prevalence threshold,

$$\frac{\sqrt{\text{True Positive Rate}(-\text{True Negative Rate} + 1)} + \text{True Negative Rate} + 1}{\text{True Postive Rate} + \text{True Negative Rate} - 1}$$

, represents local extrema such that values larger than it will only reduce the positive predictive value relative the prevalence.[16] The prevalence threshold in the minimum bounded case was 0.5, whereas it was 0 in the maximum bounded case. Comparing back to the prevalence scores, the 97.8% prevalence is beyond this threshold in the minimum bounded case. This suggests that the positive predictive value was lower than it could have been. In the minimum bounded case these values were equal, suggesting the model had relatively optimal positive predictive value. This goes back the fact that this score is partly a function of the true negative rate which differed in the two interpretations.

The threat score, $\frac{\sum \text{True Positive}}{\sum \text{True Positive} + \sum \text{False Negative} + \sum \text{False Positive}}$, represents the conditional probability that a rule holds for mice and humans given that the rule that it isn't the case that the rule doesn't work for either of humans and mice. This score was 97.7% in both bounded cases, suggesting true negatives aside, mouse rules are likely to predict human rules.

The balanced accuracy, $\frac{\text{True Positive Rate} + \text{True Negative Rate}}{2}$, is simply the arithmetic mean of the true positive rate and the true negative rate. In the minimum bounded case this was 0.5, and in the maximum bounded case it was one. Thus, including the full set of conceivable assocation rules generated by the set of genes in question would suggest perfect balanced accuracy, while considering only the extracted rules indicates being accurate 50% of the time.

The Matthews correlation coefficient is a measure of association between the association rules of the mice and humans, which is a function of all four of the true positive, true negative, false positive, and false negative fields of the confusion matrix.[17] Therefore it will be interpretation dependent as discussed above. In the minimum bounded case it is only -0.004, suggesting very little association. This is because there is relatively little variation among the fields to attempt to explain linearly or otherwise. In the maximum bounded case an overflow error occurs, giving not definable value.

The Fowlkes Mallows index, $\sqrt{\text{Positive Predictive Value} \cdot \text{True Positive rate}}$, is the geometric mean of the positive predictive value and the true positive rate. It measures the extent to which the mouse rules predict the human rules in a way that is weighted by how often the rules of humans and mice both hold. In this study it was 98.8% in both cases, suggesting that using mice inductive rules to predict human inductive rules is reliable and frequent.

The bookmaker informedness, $\text{Sensitivity} + \text{Specificity} - 1$, is a measure of the performance of using mouse inductive rules to predict human inductive rules. The smaller it is, the less reliable the inferences between mice and humans will be. The score in the minimum bounded case is effectively zero[1], suggesting low performance. The maximum bounded case suggests very high performance.

The markedness, $\text{Positive Predictive Value} + \text{Negative Predictive Value} - 1$, is a measure of divergence between the rules of humans and mice. The markedness in the minimum bounded case was low, and it was high in the maximum bounded case.

The Fisher's exact test can be applied to a confusion matrix because the confusion matrix can be interpreted as a 2x2 contingency table. This test would evaluate whether the entries of the matrix are stastically different. However, the hypergeometric distribution did not allow for the true negative values to be either zero or extremely large, and therefore it could not be evaluated in this study.

[1] The non-zero value obtained is likely due to floating point error.

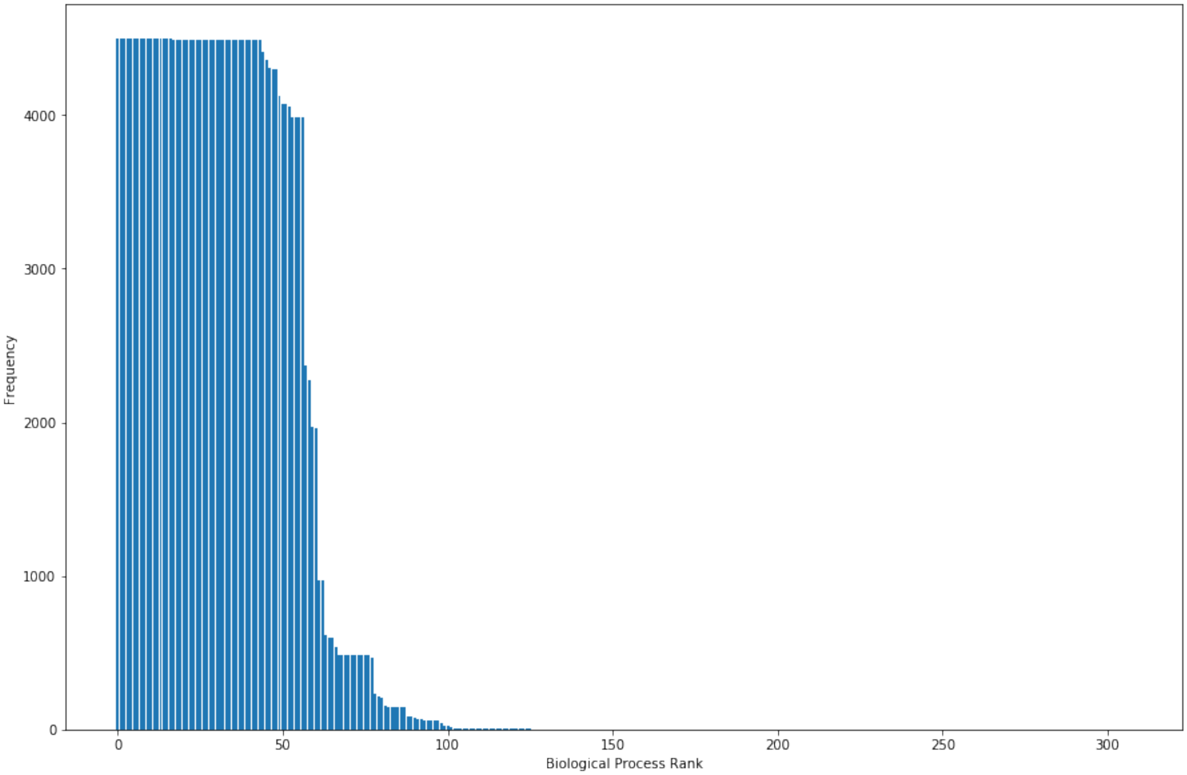


Fig. 2. Bar plot showing all the biological processes included the analysis rank ordered by the prvalence of association rules in both humans and mice.

Figure 2 shows the distribution of how frequently association rules were found to apply in each biological process, and was rank ordered for interpretability. It is clear from Figure 2 that this distribution is uneven, and that a relative minority of biological processes can have inductive rules applied to them. There is a marked dropoff after the top 40-60 processes, with the first notable drop being for those biological processes that were described by less than ~4000 association rules.

Table II gives the identity of each of the biological processeses that were described by 4000 or more association rules in both humans and mice, along with the exact number of rules. There are two many biological processes to be discussed in detail, but an attempt is made here to pick out observations and trends in the this top list of biological processes for which humans and mice share inductive rules.

Many of the biological processes in Table II are diseases, including pathways in cancer, choline metabolism in cancer, colorectal cancer, hepatitis B, human immunodeficiency virus 1 infection, Kaposi sarcoma-associated herpesvirus infection, human cytomegalovirus infection, human papillomavirus infection, proteoglycans in cancer, acute myeloid leukemia, breast cancer, central carbon metabolism in cancer, chronic myeloid leukemia, endometrial cancer, gastric cancer, glioma, hepatitis C, hepatocellular carcinoma, melanoma, non-small cell lung cancer, prostate cancer, renal cell carcinoma, human T-cell leukemia virus 1 infection, pancreatic cancer, and influenza A. A couple of observations come out of examining this list of diseases. The first that many of these are forms of cancer, which may be due to publication biases toward studying cancer. The second observation is that some of the diseases have the term ”human” in them, which is dubious when considering that these are biological processes listed for both humans and *mice*. This second observation is an apparent contradiction. Perhaps it is a mislabelling, or perhaps some of the data presumes the representativeness of mice.

The other common type of biological process found in Table II pertain to signalling systems, including the insulin signaling pathway, Ras signaling pathway, relaxin signaling pathway, Erbb signaling pathway, Fc Epsilon Ri signaling pathway, Foxo signaling pathway, neurotrophin signaling pathway, prolactin signaling pathway, sphingolipid signaling pathway, estrogen signaling pathway, thyroid hormone signaling pathway, B cell receptor signaling pathway, Pi3K-Akt signaling pathway, phospholipase D signaling pathway, Rap1 signaling pathway, signaling pathways regulating pluripotency of stem cells, T cell receptor signaling pathway, Vegf signaling pathway, Mtor signaling pathway, chemokine signaling pathway, and the camp signaling pathway. Biological signaling is at the core of organism-wide regulation of energy and matter conversion and transport, which consequently changes physiological responses at the cell and tissue levels that propogate to every functional and structural system within an organism.

TABLE II
BIOLOGICAL PROCESSES WITH THE TOP PREVALENCE OF ASSOCIATION RULES IN BOTH HUMANS AND MICE

| Biological Process | Rule Count |
|---|---|
| Autophagy | 4496 |
| Insulin Signaling Pathway | 4496 |
| Pathways In Cancer | 4496 |
| Ras Signaling Pathway | 4496 |
| Relaxin Signaling Pathway | 4496 |
| Apoptosis | 4495 |
| Choline Metabolism In Cancer | 4495 |
| Colorectal Cancer | 4495 |
| Erbb Signaling Pathway | 4495 |
| Fc Epsilon Ri Signaling Pathway | 4495 |
| Foxo Signaling Pathway | 4495 |
| Hepatitis B | 4495 |
| Human Immunodeficiency Virus 1 Infection | 4495 |
| Kaposi Sarcoma-Associated Herpesvirus Infection | 4495 |
| Neurotrophin Signaling Pathway | 4495 |
| Prolactin Signaling Pathway | 4495 |
| Sphingolipid Signaling Pathway | 4495 |
| Estrogen Signaling Pathway | 4492 |
| Human Cytomegalovirus Infection | 4492 |
| Human Papillomavirus Infection | 4492 |
| Proteoglycans In Cancer | 4492 |
| Thyroid Hormone Signaling Pathway | 4492 |
| Acute Myeloid Leukemia | 4491 |
| B Cell Receptor Signaling Pathway | 4491 |
| Breast Cancer | 4491 |
| Cellular Senescence | 4491 |
| Central Carbon Metabolism In Cancer | 4491 |
| Chronic Myeloid Leukemia | 4491 |
| Endometrial Cancer | 4491 |
| Gastric Cancer | 4491 |
| Glioma | 4491 |
| Hepatitis C | 4491 |
| Hepatocellular Carcinoma | 4491 |
| Melanoma | 4491 |
| Non-Small Cell Lung Cancer | 4491 |
| Pi3K-Akt Signaling Pathway | 4491 |
| Phospholipase D Signaling Pathway | 4491 |
| Prostate Cancer | 4491 |
| Rap1 Signaling Pathway | 4491 |
| Renal Cell Carcinoma | 4491 |
| Signaling Pathways Regulating Pluripotency Of Stem Cells | 4491 |
| T Cell Receptor Signaling Pathway | 4491 |
| Vegf Signaling Pathway | 4491 |
| Mtor Signaling Pathway | 4491 |
| Chemokine Signaling Pathway | 4413 |
| Human T-Cell Leukemia Virus 1 Infection | 4358 |
| Cholinergic Synapse | 4302 |
| Progesterone-Mediated Oocyte Maturation | 4300 |
| Pancreatic Cancer | 4299 |
| Focal Adhesion | 4122 |
| Camp Signaling Pathway | 4072 |
| Influenza A | 4071 |
| Fc Gamma R-Mediated Phagocytosis | 4057 |

### B. *Further Exploratory Data Analysis*

The section indulges in some additional exploratory data analysis to understand the dataset itself.

*1) Principal Component Analysis:* In this study the transactions where turned into a vector space by one-hot encoding the itemsets. This vector representation was then visualized in a lower-dimensional projection using principal component analysis. Principal component analysis is an orthogonal rotation technique that involves transforming the data into a new basis where the new ordered axes maximize the variation they explain. The underlying mathematics for finding the appropriate change in basis involves calculating either the eigenvalues or singular values[2] with their corresponding vectors. A subset of these vectors are selected to form a valid basis. Often principal component analysis is also considered a dimensionality reduction technique because a subset of basis vectors can be chosen after the rotation, which provides a lower-dimension representation (or lossy compression) of the original dataset.

Figure 3 illustrates a principal component analysis of the human and mouse transaction databases.[3] The top two panels show the Scree plots showing how the explained variance in the data changes as the number of top-predicting eigenvalues are included in the final basis. These plots are often used in machine learning projects to decide the number of components that should be kept in order to keep a given percentage of the original variation. Removing dimensions is desirable to avoid the curse of dimensionality and the overfitting that it often brings against losing variation in the data that might useful in prediction. What these Scree plots describe in our analysis is that relatively few of the biological processes explain most of the variation, with up to nearly 7 percent of the variation being explained in the first component with a sharp drop off afterward for both species The lower two panels show the projections of the dataset into only the top two dimensions to visually assess whether there is noticeable structure in the data. For both species there seems to be a denser region between -1 to 2.5 along the first principal component[4], with decreasing density for principcal component 1 beyond 2.5. There appears to be an outlier along the second principal component for mice that is not apparent in the human projection. I am going to make a speculation about the structure in these two scatterplots that could easily be due to noise or other factors, and requires a subjective eye of the data. To me, it appears that in both plots there appears to be a low density region within the data that splits the data into two 'arms' that extend from the left out to the right. Furthermore, the human PCA scatterplot looks similar to a reflection of the mouse PCA scatterplot roughly about the PC2=0 line. Perhaps further analysis of the eigenvalues themselves may show that there is a symmetric rotation behind the subjective similarity.

To summarize the PCA results, it appears that each single eigenvalue explains very little of the variation in each dataset, that most of the data appears to be within a single group in the 2D projection, and that there may be some symmetrical properties between the two rotated datasets.

*2) Network Visualization:* Network analysis is a large subject unto itself, with many different measures, metrics, and representations. In this study we created and visualized a graph representation of the databases for the gene enrichment within biological processes of humans and mice. Each node of these networks is a gene which appears in one or more of the biological processes, and each edge exists between two genes if they both participate in at least one biological process together. These networks are visualized in Figure 4. The network for the human gene enrichment contained 7802 nodes with 1412587 edges, giving an average degree

[2] These are technically different, but very similar in practice.

[3] Reminder: The transactions here are biological processes, and the itemsets are genes that participate in those biological processes.

[4] Note that the first component of one species is not interchangeable for the first in general beyond being the axis that linearly explains the greatest variation in the original data.

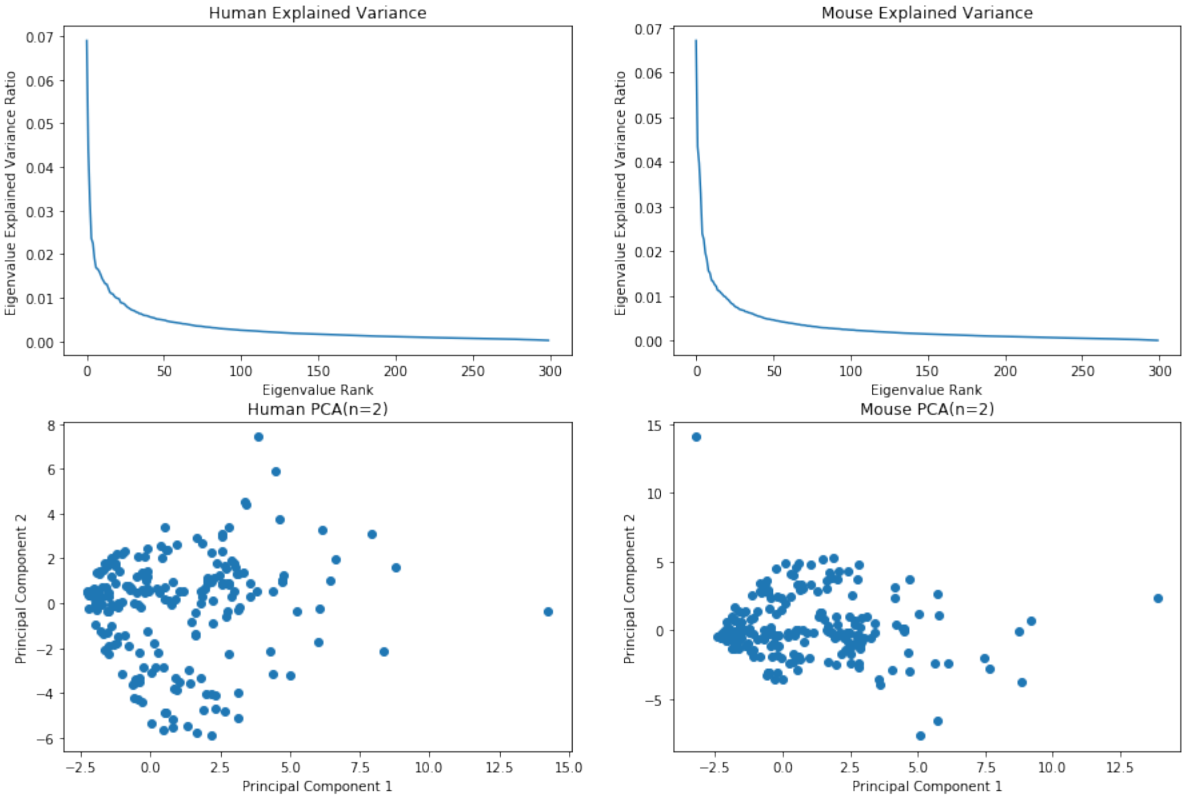
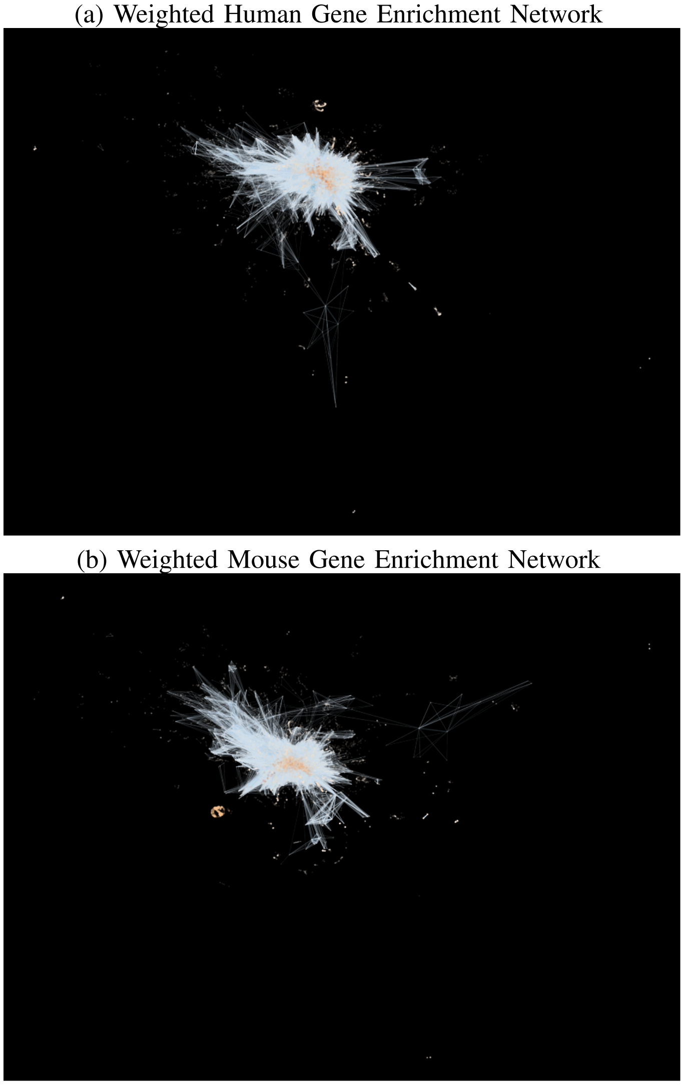


Fig. 3. Principal component analysis of one-hot encoded transaction databases. (top left) Scree plot of eigenvalues for human transactions. (top right) Scree plot of eigenvalues for mice transactions. (bottom left) Principal component plot of human transactions with two components. (bottom right) Principal component plot of mice transactions with two components.

of 362 and a graph density of 4.6%. The network for the mouse gene enrichment contained 8551 nodes with 1991850 edges, giving an average degree of 466 and a graph density of 5.4%. Despite the low graph densities, both networks were mostly connected because a minority of genes participate in most biological processes.[5] Three major features were observed about the appearance of both the human and mouse networks. The first was that most of nodes were clumped into a single major region of the spring-layout. The second was that in the the human network had a sparse 'arm' sticking down from the main "cluster" while the mouse had a similar looking arm point up-to-the-right from its main cluster. A third feature is that both networks seem to have a satellite of higher-degree nodes, with the human satellite positioned just above the main cluster and the mouse' sattelite position just to the bottom-left of its cluster. Note that the specific locations of the features are not of interest because they don't inherently capture important information about the graphs. Rather, it is the fact that they have this superficial similarities in structure that suggest there may be some agreement in structure. Checking for isomorphism of graphs in general is computationally difficult, but can be easily disconfirmed in this case by the fact that they have different numbers of nodes. Because the set differences of nodes of one graph to another are non-empty, we can conclude that there is not a subgraph of one graph that is isomorphic to second.[18] While such a task would be NP-complete, a similar notion of finding the largest subgraph of both graphs that is isomorphic would also be NP-complete. While difficult to compute, such a subgraph would identify the core commonalities in the structure of both networks.

[5] For example, amino kinases are very common because their phosphorylating action on other proteins acts as a nearly-universal form of regulation or system control.


(a) Weighted Human Gene Enrichment Network

(b) Weighted Mouse Gene Enrichment Network

Fig. 4. Weighted graphs where each node represents a gene, each edge represents existence of at least one transaction in which they participate within, and the edge weights represent the number of times that a given pair of genes occur within the same biological process. Both panels share the colour scheme where node colours range from white-to-orange as their degree centrality increases, edge colours range from white-to-blue as their weight increases, and the alpha (i.e. transparency) value was 0.1 for both nodes and edges. (a) Weighted Human Gene Enrichment Network. (b) Weighted Mouse Gene Enrichment Network.

## IV. Conclusions

This study showed that there seems to be strong agreement between humans and mice in the association rules extracted from gene enrichment data of biological processes. However, only a minority of the biological processes considered had a substantial number of association rules that applied to them.

With hundreds of thousands of association rules in common, it was not surprising that both the PCA plots and the network visualizations had common structures. Further analysis could confirm these commonalities with more objective measures, and allow for further consideration of which types of biological processes and genes tend to occur in the shared structures.

Lastly, this study identified a clear need to better understand how such databases are constructed. This was evidenced by a few observations made about the data. The first was that some of the genes listed in the mouse dataset were labeled 'human', leaving it somewhat dubious what the criterion was for including a gene in the mouse dataset. There is a legitimate concern that the construction of these gene enrichment datasets presupposed comparability of humans and mice by presuming genes confirmed in humans would also be in mice. The second issue was the prevalence of particular types of biological processes that most of the association rules occured within. While signaling processes are expected to be heavily prevalent due to their involvement in almost all metabolic pathways, the large prevalance of heavily-studied diseases such as cancer or hepatitus may be an indicate of either something biologically meaningful[6] or of a publication or annotation bias.

## References

[1] David Nelson. *Lehninger principles of biochemistry*. W.H. Freeman and Company Macmillan Higher Education, New York, NY Houndmills, Basingstoke, 2017.

[2] Why mouse matters. https://www.genome.gov/10001345/importance-of-mouse-genome. (Accessed on 10/28/2020).

[3] Niall Shanks, Ray Greek, and Jean Greek. Are animal models predictive for humans? *Philosophy, Ethics, and Humanities in Medicine*, 4(1):2, 2009.

[4] ThierryF Vandamme. Use of rodents as models of human diseases. *Journal of Pharmacy and Bioallied Sciences*, 6(1):2, 2014.

[5] M. Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000.

[6] Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Maayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, May 2016.

[7] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8(1):53–87, January 2004.

[8] Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to pythons scientific computing stack. *The Journal of Open Source Software*, 3(24), April 2018.

[9] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman,

[6] For example, it may be the case that creating uncontrolled growth (i.e. cancer) of cells can be acheived in many ways.

editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[10] Clark Glymour and Frederick Eberhardt. Hans Reichenbach. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

[11] Remco Bouckaert, Timothy G. Vaughan, Jolle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, Michael Matschiner, Fábio K. Mendes, Nicola F. Müller, Huw A. Ogilvie, Louis du Plessis, Alex Popinga, Andrew Rambaut, David Rasmussen, Igor Siveroni, Marc A. Suchard, Chieh-Hsi Wu, Dong Xie, Chi Zhang, Tanja Stadler, and Alexei J. Drummond. BEAST 2.5: An advanced software platform for bayesian evolutionary analysis. *PLOS Computational Biology*, 15(4):e1006650, April 2019.

[12] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. Algorithms for association rule mining - a general survey and comparison. *ACM SIGKDD Explorations Newsletter*, 2(1):58–64, June 2000.

[13] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD 93*. ACM Press, 1993.

[14] Pang Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson Addison Wesley, Boston, 2005.

[15] Ethan Zhang and Yi Zhang. F-measure. In *Encyclopedia of Database Systems*, pages 1147–1147. Springer US, 2009.

[16] Jacques Balayla. Prevalence threshold and the geometry of screening curves, 2020.

[17] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), January 2020.

[18] Stephen A. Cook. The complexity of theorem-proving procedures. In *Proceedings of the third annual ACM symposium on Theory of computing - STOC '71*. ACM Press, 1971.