BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 19 2010, pages 2416–2423
doi:10.1093/bioinformatics/btq451

Gene expression

Advance Access publication August 11, 2010

Cross-species queries of large gene expression databases

Hai-Son Le[1], Zoltán N. Oltvai[2] and Ziv Bar-Joseph[1,3,*]

[1]Machine Learning Department, Carnegie Mellon University, [2]Department of Pathology, University of Pittsburgh Medical School and [3]Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Expression databases, including the Gene Expression Omnibus and ArrayExpress, have experienced significant growth over the past decade and now hold hundreds of thousands of arrays from multiple species. Since most drugs are initially tested on model organisms, the ability to compare expression experiments across species may help identify pathways that are activated in a similar way in humans and other organisms. However, while several methods exist for finding co-expressed genes in the same species as a query gene, looking at co-expression of homologs or arbitrary genes in other species is challenging. Unlike sequence, which is static, expression is dynamic and changes between tissues, conditions and time. Thus, to carry out cross-species analysis using these databases, we need methods that can match experiments in one species with experiments in another species.

Results: To facilitate queries in large databases, we developed a new method for comparing expression experiments from different species. We define a distance metric between the ranking of orthologous genes in the two species. We show how to solve an optimization problem for learning the parameters of this function using a training dataset of known similar expression experiments pairs. The function we learn outperforms previous methods and simpler rank comparison methods that have been used in the past for single species analysis. We used our method to compare millions of array pairs from mouse and human expression experiments. The resulting matches can be used to find functionally related genes, to hypothesize about biological response mechanisms and to highlight conditions and diseases that are activating similar pathways in both species.

Availability: Supporting methods, results and a Matlab implementation are available from http://sb.cs.cmu.edu/ExpQ/

Contact: zivbj@cs.cmu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 25, 2010; revised on July 9, 2010; accepted on August 4, 2010

1 INTRODUCTION

Advances in sequencing technology have led to a remarkable growth in the size of sequence databases over the past two decades. This has allowed researchers to study newly sequenced genes by utilizing knowledge about their homologs in other species (Lee *et al.*, 2007). Alignment and search methods, most notably BLAST (Altschul *et al.*, 1997), have become standard tools and are extensively used by molecular biologists. Cross-species analysis of sequence data is now a standard practice. However, similar usage of expression databases has not materialized. Expression databases, including Gene Expression Omnibus (GEO; www.ncbi.nih.gov/geo/) and ArrayExpress (www.ebi.ac.uk/Databases/microarray.html) hold hundreds of thousands of arrays from multiple species (Fig. 1). Co-expression is a powerful method for assigning new function to genes within a single species (Owen *et al.*, 2003). If we are able to identify a large set of matched expression experiments across species this method can be extended and used in a cross-species analysis setting as well. Consider a human gene with unknown function that is co-expressed (across many different conditions) with a mouse gene with known function. This information can provide useful clues about the function of the human gene. This information is also useful for identifying orthologs. If a gene has multiple homologs in another species then the homolog with the highest co-expression similarity in several conditions is likely its orthologs since they are involved in the same processes in both species.

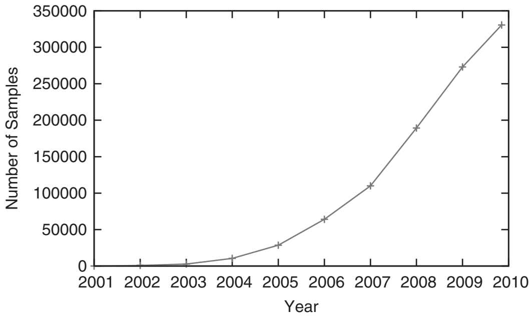

Fig. 1. Growth of microarray databases. Growth in microarray datasets deposited in GEO in the past decade. The growth resembles the impressive growth of sequence databases in the 90s.

While promising, querying expression datasets to identify co-expressed genes in other species is challenging. Unlike sequence, which is static, expression is dynamic and changes between tissues, conditions and time. Thus, a key challenge is to match experiments in one species with experiments in another species. Almost all studies that have analyzed expression datasets in multiple species relied on one of the two methods. They have either carried out experiments under the same condition in multiple species or have looked at co-expression within a species and tested whether these relationships are retained across species. Examples of the former set of methods include comparison of cell-cycle experiments across species (Jensen *et al.*, 2006), comparing response programs (Lelandais *et al.*, 2008) and comparing tissue expression between human and mouse (Su *et al.*, 2004). Examples of the latter strategy include the metaGene

*To whom correspondence should be addressed.

analysis (Stuart *et al.*, 2003) and cross-species clustering methods (Lu *et al.*, 2007). See Lu *et al.* (2009) for a recent review of these methods.

Although successful, the approaches discussed above are not appropriate for querying large databases. In almost all cases it is impossible to find a perfect match for a specific condition in the database. Even in the rare cases when such matches occur it is not clear if the same pathways are activated in the different species. For example, many drugs that work well on animal models fail when applied to humans, at least in part because of differences in the pathways involved (Bussiere, 2008). Looking at relationships within and between species would also not answer the questions we mentioned above since these require knowledge of orthologs assignment to begin with. These methods are also less appropriate for identifying one to one gene matchings because they are focusing on clusters instead.

The only previous attempt we are aware of to facilitate cross-species queries of expression data is the non-negative matrix factorization (NMF) approach presented by Tamayo *et al.* (2007). This unsupervised approach discovers a small number of metagenes (similar to principle components) that capture the invariant biological features of the dataset. The orthologs of the genes included in the metagenes are then combined in a similar way in the query species to identify related expression datasets. While the approach was successfully used to compare two specific experiments in humans and mouse, as we show in Results, the fact that the approach is unsupervised makes it less appropriate for large scale queries of expression databases.

In this article, we present a new method for identifying similar experiments in different species. Instead of relying on the description of the experiments, we develop a method to determine the similarity of expression profiles by introducing a new distance function and utilizing a group of known orthologs. Our method uses a training dataset of known similar pairs to learn the parameters for distance functions between pairs of experiments based on the rank of orthologous genes overcoming problems related to difference in noise and platforms between species. We show that the function we learn outperforms simpler rank comparison methods that have been used in the past (Fujibuchi *et al.*, 2007; Hunter *et al.*, 2001). We next use our method to compare millions of array pairs from mouse and human experiments. The resulting matches highlight conditions and diseases that are activating similar pathways in both species and can also hint at diseases where these pathways seem to differ. Given the large number of arrays in current databases our methods can also be used to aid manual annotations of cross-species similarity by focusing on a small subset of the millions of possible matches.

We note that while the discussion below focuses on microarray data and we have only tested our methods on such data, our methods are appropriate for deep sequencing expression data as well. As long as a partial orthologs list can be obtained, the methods we present below can be used to compare any expression datasets across species.

2 METHODS

2.1 Comparing microarrays across species

Our goal is to obtain a distance function that given two microarray datasets outputs a small distance between experiments that are very similar and a large distance for those pairs that study different processes or in which different pathways are activated in the two species being compared. Since we are comparing experiments from different platforms and species the first decision we made was to compare the ranking of the genes in each array rather than their expression levels [previous methods for comparing experiments in the same species have relied on ranking as well (Fujibuchi *et al.*, 2007)]. There are a number of other properties that we seek for such scoring functions. First, they should of course be able to separate similar pairs from non-similar pairs. In addition, it would be useful if the function is a metric or a pseudometric [a pseudometric satisfies all properties of a metric except for the identity, i.e. $d(x,y)$ could be 0 even if $x \neq y$]. This will guarantee useful distance properties including symmetry and triangle inequality (see Supporting Methods in Supplementary Material for the complete list). Finally, we would like to be able to determine some statistical properties for these scoring methods in order to determine a P-value for the similarity/difference between the experiments being compared (Section 2.3.1).

2.1.1 Notations We first provide notations that are used in the rest of the article. As mentioned above our function would be constructed from metrics on permutations (ordering) of ranks. Each microarray experiment is a vector in R^n, where each dimension is the expression value for a specific gene. We consider the problem of comparing a microrarray $\mathcal{X}$ of a species A with n_A genes and a microarray $\mathcal{Y}$ of a species B with n_B genes. There are m orthologs between the two species. In other words, there is a one-to-one mapping O from m species A genes to m species B genes. $1,\ldots,m$ are the orthologs, $X=\{X_i : 1 \leq i \leq m\}$ and $Y=\{Y_i : 1 \leq i \leq m\}$ are the expression values of the orthologs in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let π,σ be the rank orderings of the expression values of the orthologs in X and Y. For simplicity, we assume that there are no ties in rankings. Therefore, π,σ are two elements of the permutation group G_m. Recall that $\pi,\sigma : \{1,\ldots,m\} \rightarrow \{1,\ldots,m\}$ are bijections: $\pi(i),\sigma(i)$ are the ranks given to the ortholog i, with lowered numbered ranks given to higher expression values. Also, let I_m be the identity permutation in G_m. Finally, $\mathrm{tr}(M)$ is the trace of a matrix M.

Assume we have a metric d on G_m. For our significance analysis, we test the null hypothesis H_0 that π and σ are not associated versus the alternate hypothesis that they are. One way is to ask how large $d(\pi,\sigma)$ would be if σ were chosen uniformly at random. More formally, let D_d be the distribution of $d(\pi,\sigma)$ when σ is drawn uniformly from G_m. We reject the null hypothesis H_0 if $d(\pi,\sigma)$ is significantly smaller than $E(D_d)$. This setting is a standard approach in literature (Diaconis, 1988) (see also Supporting Fig. 1 in Supplementary Material).

2.2 Fixed distance function: Spearman's rank correlation

Below we discuss distance functions that satisfy the requirements mentioned above for cross-species analysis. We first discuss a method that does not require any parameter tuning. Such methods have been extensively used for comparing permutations. However, as we show in Section 3.2 they are less appropriate for gene expression data due to the unique properties of such data. In the next section, we discuss modification of these methods that are more appropriate for the expression data we are working with.

The Spearman's rank correlation R metric is defined as

$$R(\pi,\sigma) = \sqrt{\sum_{i=1}^{m} \big(\pi(i) - \sigma(i)\big)^2} \qquad (1)$$

In other words, it is the L_2 distance between π and σ. Hence, it is a metric. Moreover, using Hoeffding's central limit theorem it can be proved that R^2 has a limiting normal distribution (Diaconis, 1988). Note that frequently, R is standardized to have values in $[-1,1]$. This yields the widely used Spearman's rank correlation ρ.

$$\rho = 1 - \frac{6R^2(\pi,\sigma)}{(m^3 - m)} \qquad (2)$$

2.3 Adaptive Metrics

While fixed methods that do not require parameter tuning have proven useful for many cases they are less appropriate for the expression data. In such data, the importance of the ranking is not uniform. In other words, genes that are expressed at very high or very low levels compared to baseline may be very informative, whereas the exact ranking of genes that are expressed at baseline levels may be much less important. Thus, rank differences for genes in the middle of the rankings are more likely due to noise. An appropriate way to weight the differences between the rankings may lead to a better distance function between arrays. The key challenge is to determine what are the important ranks and how they should be weighted. Below we present a number of adaptive methods that can address this issue. The methods we present differ in the number of parameters that needs to be learned, and thus each may be appropriate for different cases depending on the amount of training data that exists.

2.3.1 Weighted Rank Metric Using a weight vector w of length m, we can modify the Spearman's rank correlation and define the following metric:

$$d(\pi,\sigma) = \sqrt{\sum_{i=1}^{m} \left(w(\pi(i)) - w(\sigma(i))\right)^2} \tag{3}$$

The vector w defines the weight of each rank, and thus captures the significance of each rank in measuring the association of two microarrays. Consider two arrays $(1,2,3,4)$ and $(1,3,2,4)$. Their Spearman's R distance is $\sqrt{2}$ while for a weight vector $w = (1,0,0,1)$, their distance would be 0. Such a weight vector places the weight on the top and bottom matches and disregards middle orderings.

The resulting function is no longer a metric, but rather a pseudometric in the original π, σ space ($d(\pi,\sigma) = 0$ does not imply $\pi = \sigma$). However, it is easy to see that it is a metric in the transformed $w(.)$-space because it is a L_2 distance between the vectors $w(\pi)$ and $w(\sigma)$, where $w(\pi) = \left(w(\pi_1),\ldots,w(\pi_m)\right)$ and similarly for $w(\sigma)$. In other words, the w-transformation makes some of the permutations indistinguishable indicating that the changes made are not significant, and so the two permutations result in the same weighted vector. However, for those permutations that are still distinguishable following the w-transformation the metric properties are preserved. The distribution D_d of $d(\pi,\sigma)$ when σ is drawn uniformly from G_m is asymptotically normal. See Supporting Methods in Supplementary Material for proof. We can calculate the mean and variance of D_d through exact calculation or random sampling. P-value can then be calculated based on this normal distribution.

A specific assignment of weights which is in line with our assumptions regarding the importance of genes expression ranks is the following modified Spearman's rank correlation.

2.3.2 Top–bottom R For any $0 < k < 1$ and $r > 0$, we can define w as following:

$$w(i) = \begin{cases} r(i - km) & \text{if } 1 \le i < km, \\ r(i - (1-k)m) & \text{if } (1-k)m < i \le m, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Note that genes expressed at a high level will have negative weights and those with low levels positive weights allowing the method to penalize experiments, in which genes move from one extreme to the other. All middle ranks $[km,(1-k)m]$ are assigned the same weight so genes that have ranks changed within this interval do not affect the distance at all. At the same time, it scales the high and low ranks r times to a wider range to increase the granularity of rank difference. Choosing the value of k and r can either be done using cross-validation or it could be manually specified.

2.3.3 Learning a complete weight vector w While the above method leads to different weights for different rankings it specifies a very strict cutoff, which may not accurately represent the importance of the differences in ranking. An alternative approach is to assign weights that are continuously changing based on the ranking by learning a weight vector from the training data. Here, we assume that we have access to such training data, which is indeed the case for a number of pairs of species (most notably tissue data for human and mouse as we use in Section 3). Assume we have M microarrays of species A and N microarrays of species B and for each microarray, let $\mathcal{S}$ be the set of pairs of similar arrays and $\mathcal{D}$ is the set of pairs of dissimilar arrays. If the dissimilar arrays are not known, we can select $\mathcal{D}$ as the set of all pairs that are not in $\mathcal{S}$.

Each permutation π can be represented as a binary $m \times m$ matrix M_π.

$$M_\pi(i,j) = \begin{cases} 1 & \text{if } \pi(i) = j, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Using this notation, we can define an L_2 metric d as:

$$d(\pi,\sigma) = \|M_\pi w - M_\sigma w\|_2 \tag{6}$$

$$= \sqrt{w^T \left(M_\pi - M_\sigma\right)^T \left(M_\pi - M_\sigma\right) w} \tag{7}$$

Our goal is to learn a vector w such that this distance be small for the positive set and large for the negative set. This leads to the following optimization problem:

$$\min \sum_{(x,y)\in\mathcal{S}} w^T \left(M_{\pi_x} - M_{\pi_y}\right)^T \left(M_{\pi_x} - M_{\pi_y}\right) w \tag{8}$$

$$\text{s.t} \sum_{(x,y)\in\mathcal{D}} w^T \left(M_{\pi_x} - M_{\pi_y}\right)^T \left(M_{\pi_x} - M_{\pi_y}\right) w = 1 \tag{9}$$

Note that the summation is on different groups. The optimization (top) is summed over the similar pairs, whereas the constraint (bottom) is summed over the dissimilar pair. The choice of the constant 1 on the right-hand side of (9) is arbitrary. However, replacing it with any constant $c > 0$ results only in w being multiplied by $\sqrt{c}$ that leads to the same order of scores for microarray pairs and so does not change our results. We can further simplify the problem to

$$\min w^T Z_{\mathcal{S}} w \tag{10}$$

$$\text{s.t} w^T Z_{\mathcal{D}} w = 1 \tag{11}$$

with $Z_{\mathcal{S}} = \sum_{(x,y)\in\mathcal{S}} \left(M_{\pi_x} - M_{\pi_y}\right)^T \left(M_{\pi_x} - M_{\pi_y}\right)$ and $Z_{\mathcal{D}} = \sum_{(x,y)\in\mathcal{D}} \left(M_{\pi_x} - M_{\pi_y}\right)^T \left(M_{\pi_x} - M_{\pi_y}\right)$. The matrices $Z_{\mathcal{S}}$ and $Z_{\mathcal{D}}$ are positive semi-definite since they are sums of positive semi-definite matrices $\left(M_{\pi_x} - M_{\pi_y}\right)^T \left(M_{\pi_x} - M_{\pi_y}\right)$. Although this optimization is not convex, there exists global minima based on the reformulation of this problem to finding eigenvalues of the Rayleigh quotient. The derivation is similar to Fisher's Linear Discriminant Analysis (Hastie *et al.*, 2009).

2.3.4 Relational Weighted Rank Metric A drawback of the weight vector distance metric discussed above is that it assigns weights to ranks in each microarray independent of the ranks in the other microarray. To overcome this problem, we extend the vector weight w into a full matrix W to incorporate the dependence between ranks in two microarrays. For a pair of microarrays with ortholog rankings π and σ, define a symmetric $m \times m$ matrix $M^F_{\pi,\sigma}$, whose entries (i,j) are non-zeros if and only if there exists a gene g such that g is ranked i and j in the microarrays, respectively. Formally,

$$M^F_{\pi,\sigma}(i,j) = \mathbf{1}\left[\pi^{-1}(i) = \sigma^{-1}(j)\right] + \mathbf{1}\left[\pi^{-1}(j) = \sigma^{-1}(i)\right] \tag{12}$$

In other words, $M^F_{\pi,\sigma}$ is a matrix where an entry of 1 in location (i,j) indicates that the gene in location i in the first experiment is the same as the gene in location j in the second or vice versa. By definition, $M^F_{\pi,\sigma}$ is a symmetric matrix. Note that this definition implies that if a gene g is ranked i-th in both π and σ then $M^F_{i,i} = 2$ and when $\pi = \sigma$, $M^F = 2I$. Let W be a positive semi-definite $m \times m$ matrix, with each entry $W_{i,j}$ being the weight assigned to a gene having rank i and j in the two microarrays. The larger the entries are, the more dependent the two ranks are.

Given these notations, we define the distance between the two microarrays as

$$d(\pi,\sigma)=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\left(2I-M^F_{\pi,\sigma}\right)\circ W\right)_{i,j}} \tag{13}$$

$$=\sqrt{\sum_{\substack{i,j:\pi^{-1}(i)=\sigma^{-1}(j)\\ \text{or } \pi^{-1}(j)=\sigma^{-1}(i)}}\left(\frac{W_{i,i}+W_{j,j}}{2}-W_{i,j}\right)} \tag{14}$$

$$d(\pi,\sigma)=\sqrt{\mathrm{tr}\left(\left(2I-M^F_{\pi,\sigma}\right)W\right)} \tag{15}$$

where $\circ$ is the Hadamard, or simply entry-wise product. As mentioned above, if the two permutations are identical then $M^F=2I$ and the distance is 0. Otherwise, the penalty for a disagreement of a pair (i,j) between the rankings is $\left(W_{i,i}+W_{j,j}\right)/2-W_{i,j}$. This captures both the importance of the individual ranks (very high or very low ranking genes maybe more important than middle genes) as well as the penalty for the disagreement between the pair. Equation (14) also shows that the entity under the square root is non-negative since for a positive semi-definite matrix W, $\left(W_{i,i}+W_{j,j}\right)/2\geq W_{i,j},\forall i,j$. Equation (15) follows from Equation (13) since M^F has only one entry in each column/row. In Supporting Methods (Supplementary Material), we prove that this distance function is a pseudometric in the original permutation space and a metric in the W-transformed space.

Learning algorithm: to determine the values of W using the training data, we solve the following optimization problem:

$$\min \sum_{(x,y)\in\mathcal{S}} \mathrm{tr}\left(\left(2I-M^F_{\pi_x,\pi_y}\right)W\right) \tag{16}$$

$$\text{subject to} \sum_{(x,y)\in\mathcal{D}} \mathrm{tr}\left(\left(2I-M^F_{\pi_x,\pi_y}\right)W\right)=1 \tag{17}$$

$$W\succeq 0 \tag{18}$$

Like for the weight vector the constraint (equality to 1) is arbitrary and guarantees that dissimilar arrays are distant from each other. This optimization is a semi-definite program (Nocedal and Wright, 2006). The objective function is a summation of traces of semi-definite matrices and so this is a convex optimization problem and there exists a global minimum solution. However, the matrix W is very large (m by m) and would require large amounts of training data for learning. Since such data is limited using a full rank matrix will likely lead to overfitting. Instead we seek a low-rank approximation of W. Let Z be the rank k approximation of W: $W\approx Z=YY^T$, where $Y\in R^{n\times k}$. Given these changes the optimization problem is

$$\min \mathrm{tr}\left(Y^T Z_{\mathcal{S}} Y\right) \tag{19}$$

$$\text{subject to } \mathrm{tr}\left(Y^T Z_{\mathcal{D}} Y\right)=1 \tag{20}$$

with $Z_{\mathcal{S}}=\sum_{(x,y)\in\mathcal{S}}\left(M_{\pi_x}-M_{\pi_y}\right)^T\left(M_{\pi_x}-M_{\pi_y}\right)$ and $Z_{\mathcal{D}}=\sum_{(x,y)\in\mathcal{D}}\left(M_{\pi_x}-M_{\pi_y}\right)^T\left(M_{\pi_x}-M_{\pi_y}\right)$. See Supporting Methods in Supplementary Material for a discussion on how to further regularize this optimization problem and how to solve it using augmented Lagrangian approach.

3 EXPERIMENTS AND RESULTS

We first used a training dataset from human and mouse tissues to learn parameters for our distance functions and to test the different methods on a dataset for which the correct answer is known. We next downloaded a large number of microarray expression datasets from GEO and applied our distance function to select pairs of experiments that are similar. For this section, we consider the cross-species analysis between human (*Homosapiens*) and mouse (*Musmusculus*) biological samples. We obtained the list of 16 376 human and mouse orthologs from Inparanoid (inparanoid.sbc.su.se).

3.1 Gene variance

Although the methods described above can work for any number of orthologs, the larger the number the more data we would need to fit the weight vector and matrix methods. Since all our expression levels were log ratios to a reference data (see below), we have excluded from the analysis genes that did not vary much *within* each species. We selected the top 500 most varying orthologs for further analysis. We note two things. First, methods that are not affected by overfitting (in our case Spearman's correlation and TBR) were also tested using all orthologs with results very similar to the results obtained from the 500 gene list. Second, while such a selection favors genes with high variance across a large number of experiments, at no stage in the selection have we considered the agreement between the actual levels of orthologous genes in specific experiments.

3.2 Testing distance metrics on data from human and mouse tissues

For evaluation and comparisons of all metrics discussed in this article, we used an expression dataset, which we call 'Toronto dataset', consisting of expression profiles for 26 human tissues and their corresponding tissues in mice (Chan *et al.*, 2009). These 26 tissues pairs were profiled using species-specific custom arrays. For each tissue, we had one human and one mouse arrays, which were processed and normalized by the authors of Chan *et al.* (2009). See Supporting Table 1 in Supplementary Material and http://sb.cs.cmu.edu/ExpQ/ for the list of tissues.

We used 2-fold cross-validation with 10 random permutations of tissues to compare the performance of the NMF method (Tamayo *et al.*, 2007) and the five different distance metrics discussed above. For Pearson's correlation, we select the varying 500 genes based on their expression values. For NMF, we used the R code provided by the authors, which also performs model selection to limit the number of metagenes (Brunet *et al.*, 2004). The human samples were used to discover the metagenes and the mouse orthologs of these genes were used for the mouse metagenes. For training of the methods discussed in this article, we use the set of similar tissues as the positive set and all the remaining pairs as negative examples. Using parameters learned in the training phase, we rank all test pairs by their distance and plot a Precision–Recall (PR) curve for all methods. Since the dataset is highly skewed (i.e. there are many more negative than positive pairs), PR curves provide a more informative picture of the metrics' performance than the receiver operator characteristic (ROC) curves (Davis and Goadrich, 2006).

3.2.1 Comparison of cross-species comparison metrics As can be seen in Fig. 2 most methods (except for Spearman's rank correlation) achieved a very high precision to begin with (80% and higher). However, this precision level drops and when reaching 20% recall only the weight matrix method achieves a precision that is >90%. Since there are hundreds of thousands of expression experiments in GEO, precision is more important than recall for our goals. At these high precision rates, the weight matrix method dominates the other methods we have considered and thus we used it in all subsequent analysis.

As for the other methods, we believe that Spearman's rank correlation performs worse than Pearson's correlation because the test dataset is well normalized so non-parametric methods loose statistical power. However, in application to large, heterogenous,

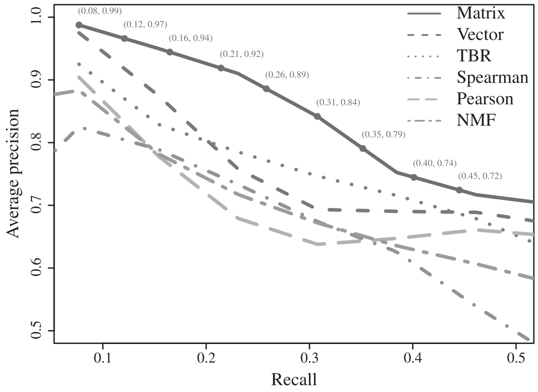

Fig. 2. Comparison of different metrics using human–mouse tissues. PR curves of Spearman's rank correlation, TBR, NMF, Vector and Matrix Weight metrics.

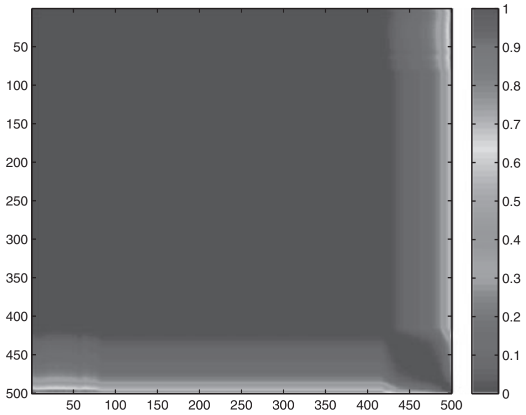

Fig. 3. The penalty for a disagreement of a pair (i, j) between the rankings, $\left(W_{i,i}+W_{j,j}\right)/2-W_{i,j}$ as shown in (14), learned from the human and mouse tissues data.

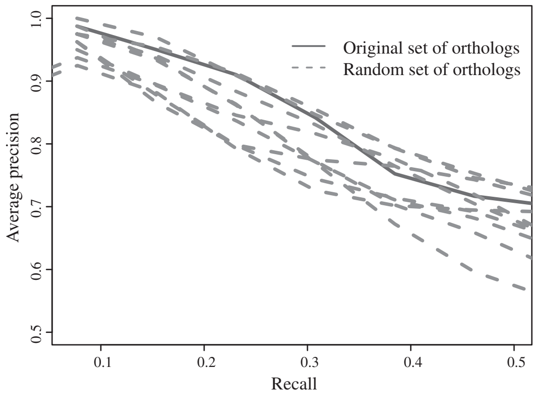

Fig. 4. PR curves for the Matrix Weight metric when starting with fewer orthologs. The solid curve is the result when starting with all orthologs (same curve as in Fig. 2).

datasets the assumption of normalization across the datasets is less likely. For NMF, the fact that it is unsupervised and does not use information from the query species to construct the components likely led to its weaker performance. The results presented in Figure 2 used an approximation matrix with Rank 3. We have also tested other ranks (recall that Rank 1 is the weight vector shown on the figure as well). We observe that both Ranks 2 and 4 do not improve the overall success (http://sb.cs.cmu.edu/ExpQ/), and so we have focused on Rank 3 matrices for the reminder of this article.

We have repeated the above analysis (comparison of methods) using another, independent, human–mouse tissue dataset, which we term the 'Novartis dataset', from Su *et al.* (2004). As we discuss in Supporting Results in Supplementary Material, this additional analysis agrees with the results presented above indicating that our method is robust to the specific data used and to the different platforms in these two studies.

Figure 3 presents the residual weights $\left(W_{i,i}+W_{j,j}\right)/2-W_{i,j}$, which are the penalties for differences in a ranked pair as shown in (14). High (dark) values indicate bigger penalty while lower (light) values indicate that the penalty is smaller. Interestingly, the method seems to focus more on the repressed genes and puts a higher weight on genes that move from being repressed to being upregulated or at a medium expression level.

3.2.2 Effect of ortholog assignment on the performance of the Matrix method Inparanoid contains over 10 000 known orthologs between human and mouse making them one of the best annotated pairs of species. As noted above, from this set we select a subset of 500 genes and use these in our algorithms. To test whether our methods would be appropriate to other species pairs for which much fewer orthologs are known, we repeated the analysis discussed above starting with a smaller set of orthologs. We selected random sets of 2000 orthologs (roughly 12% of all orthologs), and then reran our method using this initial set (selecting the top 500 varying genes from this smaller subset and running the matrix algorithm discussed above). Figure 4 presents results for seven of these random sets. The solid curve presents the results when starting with the full set of orthologs. As can be seen our method is robust and is appropriate for pairs of species with much fewer known orthologs as well.

3.3 Identifying similar experiments in GEO

The previous section shows that our weight matrix performs better than standard metrics on the Toronto and Novartis datasets, and moreover can get a very high precision for the recall value of 20%. Our goal is to apply this new metric for retrieving cross-species similar pairs of microarray experiments in a large dataset.

3.3.1 Data collection We downloaded 715 human and 769 mouse datasets from GEO and used GDS data and metadata to identify control samples for each dataset (http://sb.cs.cmu.edu/ExpQ/). Such samples are important for properly normalizing and transforming the data so that all data used is log 2 ratio of the response sample to its control. We excluded from the analysis all datasets for which we could not positively identify the control sample leaving us with 3416 human and 2991 mouse microarrays from 535 human and 641 mouse datasets.

3.3.2 Identification of associated pairs of microarrays We used the weight matrix trained using the full set of human–mouse tissue pairs. We used the results of Figure 2 to select a similarity cutoff corresponding to the cutoff that led to 95% precision and 10% recall.

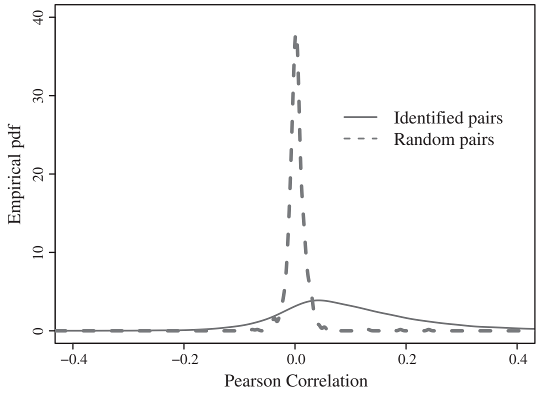

Fig. 5. Solid curve: correlation of orthologs not used for training in a random sample of 301 453 microarray pairs from human and mouse. Dashed curve: correlation of orthologs not used for training in the set of microarray pairs selected by our method.

Using this cutoff, we ended up with 301453 pairs of microarrays whose distances are smaller than the cutoff which is roughly 3% of all pairs tested. These pairs are from 14 493 dataset pairs (many array pairs are from the same pair of human and mouse datasets).

We also looked at the distribution of scores under the null hypothesis (since >95% of microarray pairs are not similar, this can be done by selecting random human–mouse array pairs) and determined that the P-value for the null hypothesis is uniformly distributed, as expected. As a sanity check for our results, we also computed the Pearson's correlation across the pairs determined to be significant by our method for all human and mouse orthologs that were not part of the 500 genes we used for learning the parameters. Figure 5 shows the histogram of this correlation and the histogram of the correlation for the same set of genes in a randomly selected set of 301453 microarray pairs. As can be seen the selected experiments are indeed more similar for many of the orthologs when compared to random selected pairs indicating that our method can identify correlated array pairs without using the experiment description.

3.3.3 Description and dataset analysis The list of pairs derived by our method allows us to address many questions. We first asked what conditions/organs/tissues are the most similar between human and mouse in terms of expression. We used the titles provided in the metadata section of the GDS to identify common words that are significantly over-represented in the microarray pairs we extracted. For each pair of similar experiments, a word that appears in both titles could provide information about the relationship between the pair. For each word, we have also computed the number of times it appeared in a title for all microarrays used from each species and the expected number of times it should have appeared in the pairs we selected. Using the hypergeometric distribution, we computed the over-representation P-value for each word. Table 1 presents the results of the analysis of over-represented words in matched titles. As can be seen some organs and tissue types are much more represented than others. For example, brain, muscles and blood appear to have similar expression patterns between the two species. Certain conditions are also over-represented, most notably immune response. Several words are associated with experiments related to such response including different types of cells participating in the response (macrophages, dendritic, cd8).

Table 1. Top 14 words identified in titles of pairs determined to be similar

Rank	P-value	Word	#Pairs	
			Identified	Expected
1	7.14429e-13	MUSCLE	121	28.46752
2	7.39409e-13	DENDRITIC	24	2.13506
3	1.76946e-11	SKELETAL	42	12.12506
4	3.12418e-11	MACROPHAGE	18	2.21414
5	1.89634e-08	ERYTHROID	6	0.15815
6	2.52933e-08	OBESITY	9	0.63261
7	8.35063e-08	HEMATOPOIETIC	13	1.84512
8	2.36749e-07	BRAIN	19	4.42828
9	1.52768e-06	CD8+	5	0.18451
10	1.67619e-06	CARDIAC	6	0.34266
11	1.45374e-05	STEM	43	20.87618
12	2.02795e-05	HAIR	5	0.31631
13	9.19217e-05	FIBROBLASTS	12	3.08398
14	2.04560e-04	AIRWAY	7	1.15979

#Pairs identified is the number of time this pair was observed. #Pairs expected is the number of time expected based on single species occurrences. The P-value is computed using the hypergeometric distribution.

In contrast, cancer, one of the most common words in the human studies (roughly 10% of human datasets contained cancer in the title) was not over-represented supporting recent results that most mice are not an ideal model system for at least some types of cancer (Sharpless and Depinho, 2006). We repeated this analysis using the abstracts provided instead of the titles leading to similar results (see http://sb.cs.cmu.edu/ExpQ/ for full results).We have also looked beyond pairwise similarities and identified entire datasets (GDS files) that contained several similar pairs of arrays between human and mouse. An expert pathologist (Oltvai) manually inspected the top 100 matched datasets and determined that over 80% of them make biological sense (see Supporting Table 2 in Supplementary Material). Many of the datasets identified as similar contained experiments for the same tissue (most notably muscle, but also blood and brain). However, some of the matches were less obvious. Fibrosis is a chronic progressive and often lethal lung disease. One of the top 50 matches in our results was between a human dataset titled non-diseased lung tissue (GDS1673) and the mouse dataset titled Pulmonary fibrosis (GDS251). However, upon a closer inspection of the mouse dataset it can be seen that it compares two mouse strains treated with bleomycin. One is determined to be susceptible to fibrosis (C57BL6/J), whereas the other is determined to be resistant (BALB/c). When looking at the similarities computed by our method it can be seen that the vast majority of the top 100 matches are for the BALB/c strains. Thus, our cross-species comparisons can be used to identify cases in which similar pathways are activated even though the conditions may be different.

3.3.4 Quarrying GEO to identify cycling mouse genes To demonstrate the utility of our method for quarrying large cross-species databases like GEO, we used a set of 50 known human cycling genes extracted from (Whitfield *et al.*, 2002). For each of these genes we used all 301 453 microarray pairs determined to be similar to identify the set of similarly expressed mouse genes using Spearman's correlations (regardless of their sequence similarity). We retrieved the top 10 most similar mouse genes for each query

Table 2. GO enrichment analysis for mouse genes using STEM

		# Genes			
Rank	Category name	Assigned	Expected	*P*	*P* adj
1	cell cycle	39.0	9.1	8.5E-15	<0.001
2	cell division	26.0	4.5	5.5E-13	<0.001
3	cell cycle phase	26.0	4.7	1.6E-12	<0.001
4	M phase	24.0	4.2	4.8E-12	<0.001
5	cell cycle process	26.0	5.5	4.6E-11	<0.001
6	mitotic cell cycle	21.0	3.8	2.4E-10	<0.001
7	mitosis	17.0	2.9	6.7E-9	<0.001
8	nuclear division	17.0	2.9	5.8E-9	<0.001
9	M phase of mitotic cycle	17.0	3.0	6.7E-9	<0.001

human gene resulting in a set of 206 genes. Note that the database we used contained a diverse set of experiments and, while a few may have been focused on cell-cycle studies the vast majority were not. Importantly, our analysis here did not rely on any specific cell-cycle time series dataset.

We used STEM (Ernst and Bar-Joseph, 2006) to determine significant GO categories associated with this list of mouse genes. As can be seen in Table 2, all top categories that are enriched for this set are related to cell cycle (including cell cycle itself). The set of mouse genes contains orthologs of the original set of human genes including CDC2A, a cell division control protein and CCNB1, an essential component of the cell-cycle regulatory machinery. The list also contains many known mouse cell-cycle genes with no homologs on the human list. These include members of a highly conserved complex, which is essential for the initiation of DNA replication (ORC1L and ORC6L), and PRIM1 and PRIM2 which are involved in chromosomal replication during cell cycle. See http://sb.cs.cmu.edu/ExpQ/ for complete list. These results highlight the potential use of our method for identifying functionally related genes across species.

4 CONCLUSIONS AND FUTURE WORK

The growth of microarray databases opens the door to applications that can simultaneously query sequence and expression databases to identify both static and dynamic matches. However, these methods would require a set of matching expression datasets in the species being queried. Such matches are hard to come by. It is rare to find the exact same experiment (condition, time, tissues, etc.) in multiple species. To allow the use of these databases, we looked at several different distance metrics between expression experiments. We defined a new distance function that utilizes the ranking of orthologs in both species. Our method uses a training dataset to learn weights for differences in rankings between the species and these differences are then summed up to determine the similarity between the two experiments. Testing this method on a training dataset of known similar pairs showed that it indeed improves upon other distance measures and that it can achieve high precision.

We used our new distance function to retrieve similar experiment pairs from GEO. The set of experiments identified by our method allowed us to look at questions regarding the conditions and tissues that activate similar expression patterns in human and mouse and to find a set of cycling mouse genes based on a set of known human cycling genes. Many of these mouse genes are known to be cycling and the rest of the genes identified are candidates for further study into their role in the cell cycle.

Our method attempts to learn a new distance function for permutations based on the training data. There has been recent work in Machine Learning on trying to learn new distance function for feature vectors (Bar-Hillel *et al.*, 2005), though we are not aware of any work so far that attempted to learn such methods for permutations. A number of the methods developed for feature vectors were later kernelized allowing for much faster computations. It would be interesting to see if the Matrix weight method discussed in this article can also be kernelized. We have primarily relied on one to one orthology matches for computing the distance between pairs of experiments. Since many orthology assignments are many to one or many to many, methods that can utilize such information may be able to improve upon the results suggested in this article. Our overall goal is to compile a large set of expression pairs that can be used for querying human and mouse genes. As we noted in the introduction our method can also help in distinguishing between orthologs and homologs by looking for genes with similar sequence that are also co-expressed in the set of similar experiments. We would also like to extend this work to other species and we are looking for training data for these species.

Funding: National Institute of Health grant 1RO1 GM085022 and the National Science Foundation grant DBI-0965316 and CAREER award 0448453 to Z.B.J. (in part).

Conlict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bar-Hillel,A. *et al.* (2005) Learning a mahalanobis metric from equivalence constraints. *J. Mach. Learn. Res.*, **6**, 937–965.

Brunet,J.P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, **101**, 4164–4169.

Bussiere,J.L. (2008) Species selection considerations for preclinical toxicology studies for biotherapeutics. *Expert Opin. Drug Metab. Toxicol.*, **4**, 871–877.

Chan,E.T. *et al.* (2009) Conservation of core gene expression in vertebrate tissues. *J. Biol.*, **8**, 33.

Davis,J. and Goadrich,M. (2006) The relationship between precision-recall and ROC curves. In *ICML'06: Proceedings of the 23rd International Conference on Machine Learning*. ACM, New York, NY, pp. 233–240.

Diaconis,P. (1988) *Group Representations in Probability and Statistics. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 11.* Institute of Mathematical Statistics, Hayward, CA.

Ernst,J. and Bar-Joseph,Z. (2006) STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, **7**, 191.

Fujibuchi,W. *et al.* (2007) CellMontage: similar expression profile search server. *Bioinformatics*, **23**, 3103–3104.

Hastie,T. *et al.* (2009) *The Elements of Statistical Learning*. Corrected edn. Springer, New York, NY.

Hunter,L. *et al.* (2001) GEST: a gene expression search tool based on a novel Bayesian similarity metric. *Bioinformatics*, **17** (Suppl. 1), S115–S122.

Jensen,L.J. *et al.* (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature*. **443**, 594–597.

Lee,D. *et al.* (2007) Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.*, **8**, 995–1005.

Lelandais,G. *et al.* (2008) Genome adaptation to chemical stress: clues from comparative transcriptomics in Saccharomyces cerevisiae and Candida glabrata. *Genome Biol.*, **9**, R164.

Lu,Y. *et al.* (2007) Cross-species microarray analysis with the OSCAR system suggests an INSR−Pax6−NQO1 neuro-protective pathway in aging and Alzheimer's disease. *Nucleic Acids Res.*, **35**, W105–W114.

Lu,Y. *et al.* (2009) Cross species analysis of microarray expression data. *Bioinformatics*, **25**, 1476–1483.

Nocedal,J. and Wright,S.J. (2006) *Numerical Optimization*. *Springer Series in Operations Research*. Springer, New York.

Owen,A.B. *et al.* (2003) A gene recommender algorithm to identify coexpressed genes in C. elegans. *Genome Res.*, **13**, 1828–1837.

Sharpless,N.E. and Depinho,R.A. (2006) The mighty mouse: genetically engineered mouse models in cancer drug development. *Nat. Rev. Drug Discov.*, **5**, 741–754.

Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.

Tamayo,P. *et al.* (2007) Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proc. Natl Acad. Sci. USA*, **104**, 5959–5964.

Whitfield,M.L. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.