BIOINFORMATICS

Vol. 28 ISMB 2012, pages i258–i264
doi:10.1093/bioinformatics/bts205

Matching experiments across species using expression values and textual information

Aaron Wise[1], Zoltán N. Oltvai[2] and Ziv Bar-Joseph[1,3,*]

[1]Lane Center for Computational Biology, Carnegie Mellon University Pittsburgh, PA, 15213, USA [2]Department of Pathology, University of Pittsburgh Medical School and Pittsburgh, PA, 15261, USA [3]Machine Learning Department, Carnegie Mellon University Pittsburgh, PA, 15213, USA

ABSTRACT

Motivation: With the vast increase in the number of gene expression datasets deposited in public databases, novel techniques are required to analyze and mine this wealth of data. Similar to the way BLAST enables cross-species comparison of sequence data, tools that enable cross-species expression comparison will allow us to better utilize these datasets: cross-species expression comparison enables us to address questions in evolution and development, and further allows the identification of disease-related genes and pathways that play similar roles in humans and model organisms. Unlike sequence, which is static, expression data changes over time and under different conditions. Thus, a prerequisite for performing cross-species analysis is the ability to match experiments across species.

Results: To enable better cross-species comparisons, we developed methods for automatically identifying pairs of similar expression datasets across species. Our method uses a co-training algorithm to combine a model of expression similarity with a model of the text which accompanies the expression experiments. The co-training method outperforms previous methods based on expression similarity alone. Using expert analysis, we show that the new matches identified by our method indeed capture biological similarities across species. We then use the matched expression pairs between human and mouse to recover known and novel cycling genes as well as to identify genes with possible involvement in diabetes. By providing the ability to identify novel candidate genes in model organisms, our method opens the door to new models for studying diseases.

Availability: Source code and supplementary information is available at: www.andrew.cmu.edu/user/aaronwis/cotrain12.

Contact: zivbj@cs.cmu.edu

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

1 INTRODUCTION

Cross-species analysis has been at the center of genomics research for decades. Some of the most influential computational biology work, including BLAST (Altschul *et al.*, 1997) and various alignment methods (Needleman and Wunsch, 1970) were aimed at comparing genomics data across species. In addition to answering several basic research questions [including issues related to evolution (Stark *et al.*, 2007) and development (Barr *et al.*, 2003)], cross-species analysis is extensively used by pharmaceutical companies. Indeed, almost all drugs are initially developed and tested using model organisms, and knowledge about the relationship between target genes in these organisms and corresponding human genes is crucial for successful drug development (Kaletta and Hengartner, 2006).

While most work to date has focused on the analysis of sequence data across species, other types of genomics data are rapidly accumulating. One of the most abundant types of genomics data are gene expression data. Unlike sequence data, expression data changes between conditions, time points and developmental stages and is thus extremely useful for studies that involve responses to various treatments. Gene expression databases from microarray studies have grown exponentially over the last decade (Le *et al.*, 2010). Other technologies, including RNA sequencing, are also generating large expression datasets in multiple species. This leads to a key challenge: How can we effectively mine these databases to identify similarities and differences in gene expression across species that complement sequence data?

A prerequisite for cross species analysis is the ability to match data in one species to data in another. This can be easily done for sequence data since DNA is context independent and the nucleotides and amino acids are universal. However, things become more challenging when using expression data. First, genes need to be matched across species, and not all orthologs are currently known. More importantly, expression data are condition specific, continuous, sometimes dynamic and often much noisier than sequence data. This makes it hard to identify experiments that can be matched to find genes that are expressed in a similar way across species.

To address this issue, several researchers performed controlled experiments in which the same biological system was studied under the same condition, in the same lab, and in multiple species. Examples include the cell cycle (Rustici *et al.*, 2004), immune response (Zinman *et al.*, 2011), various tissues (Su *et al.*, 2004), drug response (Kuo *et al.*, 2010) and development (Rifkin *et al.*, 2003). See Lu *et al.* (2009) for a recent review. While these studies successfully identified similarities and differences leading to new insights regarding conservation and response mechanisms, this success only serves to strengthen the question mentioned above: Can we develop methods to mine the vast number of expression experiments currently deposited in public databases so that they can also be used in such a cross-species analysis framework?

Relatively little work has been carried out to date to address this general question (especially when compared with work that focuses on the cross-species analysis of sequence data). One previous approach by Tamayo *et al.* (2007) used non-negative matrix factorization (NMF) to perform the unsupervised discovery of a small set of metagenes that are a linear combination of gene expression levels in one of the species being compared. By similarly combining the orthologs of these genes into metagenes in another

[*]To whom correspondence should be addressed.

© The Author(s) 2012. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

species, expression experiments were compared across species to identify matched pairs. Another method that, similar to the NMF method, only uses expression data, was proposed by Le *et al.* (2010). Using a subset of the orthologs between two species and a small training set, their method learns a new distance function between microarray experiments in the two species. That distance function is then used to select a new set of matched arrays which serve as a basis for querying gene similarities across species. Le *et al.* have shown that their method improved upon prior methods (including the NMF method). However, similar to the NMF method and other methods, the Le *et al.* method did not utilize all available information, which reduced its performance. Specifically, the method only used expression values while expression databases also provide textual information both regarding the dataset (an abstract) and regarding the individual arrays (time point, exact condition, etc.).

In this article, we extend the Le *et al.* method so that we can integrate expression values and text when searching for matched array pairs. Our new method utilizes latent semantic analysis (LSA) (Deerwester *et al.*, 1990) to match abstracts across species. Using training data, we initially learn a model for LSA [parametrized by how many dimensions to keep during singular value decomposition (SVD)] and use this model to rank a set of pairs of arrays across species. We then combine our LSA model with the expression analysis method from (Le *et al.*, 2010) using a co-training framework.

In co-training [which belongs to a larger class of semi-supervised learning methods (Chapelle *et al.*, 2006)], two models are iteratively improved by continuously increasing the training (labeled) set at each iteration based on the agreement of the models on the unlabeled examples. As we show, the resulting combined model improves upon the expression-only method. This is apparent both when using a standard train-test approach and when analyzing biological data for functional assignments. We then use our new model to identify a set of matched arrays which serves as the basis for the cross species queries. We manually analyzed the accuracy of the top set of pairs identified by the combined model concluding that our method can successfully identify the relevant matches. We then used the new matches to identify genes potentially involved in diabetes based on correlation to genes with known involvement in our matched pairs.

2 METHODS

We use a co-training approach to iteratively learn the parameters for two models of microarray similarity, each of which uses a different set of features as input. The first model is used to determine the similarity of two microarrays based on their expression values. From a set of training data, it learns a distance metric for comparing expression values of orthologs across the two species. The second model is used to determine the similarity of text (in this case, the descriptions attached to microarrays). For this model, LSA is used, which maps the text into a low-dimensional space where dimensions correspond (roughly) to semantic concepts. Texts are then determined to be similar or not in this low-dimensional space. The co-training algorithm begins by training each of the two models separately using hand-curated training data (i.e. labeled data). Then, each model is used to score the unlabeled data, finding the most similar pairs of microarrays using each of the two methods. It then finds microarray pairs which rank highly using both models. These pairs are added to the labeled list and (together with the original set of labeled data) are used as training data for the next iteration.

2.1 Gene expression comparison

The first of the two models we used in the co-training method is a distance metric that allows us to score the similarity of two microarrays from different species based on the ranks of known orthologs. We learn this distance metric in a similar manner to the one described in (Le *et al.*, 2010). The training requires a set of positive and negative examples. (Positive examples include pairs of arrays that are representing a similar condition and tissue whereas negative examples are pairs that represent different conditions/tissues.) In addition, we use a set of known gene orthologs between the two species.

To avoid issues related to different platforms and normalizations, we rely on the rank order of the genes rather than on their actual values. For each array from the two species we record the permutation induced by the rank order of the orthologs expression levels. This ranking of the log expression ratios (relative to control) is encoded using a matrix M in the following way:

$$M(i,j) = \begin{cases} 1 & \text{the rank of ortholog } i \text{ is } j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

With this definition of M, we define a distance function between two microarrays based on a weighted difference of the permutation induced by the order of the orthologs. Specifically, we set:

$$d(M_\pi, M_\sigma) = \sqrt{w^{\mathrm{T}}(M_\pi - M_\sigma)^{\mathrm{T}}(M_\pi - M_\sigma)w} \quad (2)$$

where M_π is a microarray from one species, M_σ is a microarray from the other and w is a weight vector.

Our goal is to learn a vector w that minimizes the distances on our positive training set, and maximizes distances on the negative training set. Furthermore, we look for a distance function that penalizes 'large' deviations in ranking (for example, moving from a highly expressed status to a repressed status) while at the same time allowing genes to move a few spots up or down the ranking without penalty (due to noise, a gene ranked 1000 in one experiment can be ranked 1100 in another even if its activity does not change much). Of course, manually quantifying what constitutes a 'large' deviation is very hard to do. We thus use the training data in an optimization procedure to find the correct values for w. This optimization problem is equivalent to finding eigenvalues of the Rayleigh quotient. See Le *et al.* (2010) for complete details.

Once we learn such a w, we can measure the distance between any pair of microarrays (one microarray from each species). We use this metric to rank order all array pairs between the two species being analyzed.

2.2 Textual comparison

The second model we use in the co-training procedure compares two microarrays based on the similarity of the abstract text that accompanies the gene expression dataset. In GEO (the Gene Expression Omnibus), as well as in other expression databases, an abstract is required when depositing datasets (a 'GDS' in GEO). Note, however, that datasets include multiple arrays (in many cases far more than 10). Thus, while the textual score will allow us to find similar experiments across species it may not be enough for matching individual microarrays. Thus, we need the co-training procedure which can also utilize expression values.

We use LSA to score textual similarity (Deerwester *et al.*, 1990). To prepare abstracts for scoring, we use a stemmer to remove word suffixes. Then, a blacklist of common non-content words is used to restrict abstracts to words with probable biological meaning. We build a term-document co-occurrence matrix, where an entry $N(t_i, d_j)$ is equal to the number of occurrences of term i in document j.

We then use the term frequency–inverse document frequency (TF–IDF) transformation on our co-occurrence matrix. TF–IDF weighting increases the weighting of words that are proportionally rarer in the document corpus; this is desirable because words that occur in fewer experiment descriptions are likely to be more valuable in distinguishing a given experiment from others. For example, the name of a specific gene under study is rare, and two abstracts containing the same gene are likely to be related; however the description of an experiment as a time series is more common, and proportionally less useful. More specifically, to perform the TF–IDF transform we determine

$$\mathrm{TF}(t_i,d_j)=\frac{\#\text{word } t_i \text{ in } d_j}{\#\text{words in } d_j} \tag{3}$$

$$\mathrm{IDF}(t_i,d_j)=\log\left(\frac{D}{\sum_{k=1}^{D} I(t_i\in d_k)}\right) \tag{4}$$

where D is the number of documents in our corpus. Using these values we set $N(t_i,d_j)$ to $\mathrm{TF}(t_i,d_j)*\mathrm{IDF}(t_i,d_j)$.

Finally, we perform SVD on the TF–IDF matrix to produce a low-dimensional projection of the TF–IDF matrix. Abstracts for pairs of microarrays can then be compared by taking the cosine of their projections in the lower dimensional space.

The only parameter we need to set in this procedure is the number of dimensions X to preserve when performing SVD. This number plays an important role. If X is too large (keeping many dimensions) then we may overfit leading to an inability to match correct pairs. On the other hand if X is too low (few dimensions) the resulting model may not be specific enough leading to many erroneous matches. To find the right value for X we again rely on the training data. We first compute the distance of all microarray pairs for each choice of the number of dimensions to preserve from SVD. For each dimension, we normalize the scoring of pairs so that the mean score is 0, and the SD is 1. Then, we sum the similarity scores for each pair in the positive training set. The dimensional cutoff that has the highest sum of similarity scores for the training set is chosen.

2.3 Iterative co-training

We iteratively refine our models using a co-training technique (Blum and Mitchell, 1998). Co-training, a form of semi-supervised learning, is an iterative machine learning technique that allows us to combine two models which use different, ideally independent, views (features) of the data. The core idea of co-training involves four steps:

(1) Train two models using a set of labeled training examples;

(2) Assign labels to all unlabeled examples using both models;

(3) Choose examples that were labeled the same by each model and add these to the training examples; and

(4) Repeat Steps 1–3 until convergence or a set number of iterations.

The main advantage of co-training is the fact that we can use the vast amount of unlabeled data (pairs of microarrays for which we do not know if they are similar or not) to improve our classifiers. While it is a hard manual task to actually label pairs of arrays from two species (it took one of us several hours to manually label 100 pairs), the number of unlabeled pairs is very large (roughly 10 million). Using two different views of the data allows us to use initially unlabeled microarrays as an additional source of training data for our models.

In our co-training procedure we combine models for both the expression values and the text abstract that are associated with the microarray. Each of these two factors provides a different view of microarray similarity: the expression levels give us a measure of similarity in expression response whereas the text gives us a measure of similarity in tissue and experimental manipulation. We expect that pairs that are similar in both expression level and text will be of higher quality (i.e. more similar) than pairs that score highly on just one of the two metrics.

In each iteration, we first train each model, and then evaluate all unlabeled microarray using both distance metrics (gene expression similarity and textual similarity). Then we score all microarray pairs using the trained expression and textual distance metrics.

At this point we perform the co-training step: we take all array pairs in the intersection of the top 1% of expression similarity scores and the top 1% of textual similarity scores and add them to the initial positive training set. Finally, we check for termination conditions: whether the positive set is unchanged or a certain number of iterations have occurred. When we reach the stopping criteria we use the two models that we have learned to derive a final set of matched arrays using the intersection of the top scoring pairs using each method.

2.4 Manually matching experiments for use as a training set

To obtain a training set we followed the following procedure. The abstracts of data sets (GDSes) from GEO were scored using LSA (using a default 670 dimensions), and the top 100 experiments were then manually evaluated to determine if experiment pairs are indeed similar (same tissue and condition). Individual microarray pairs were similarly manually evaluated using single array descriptors (time point, specific treatment for that sample). Following these steps we obtained a total of 138 labeled microarray pairs. These pairs were used as the initial positive training set. See website for the complete list of matched arrays used for training.

3 RESULTS

We performed experiments by searching for matches between human (*Homo sapiens*) and mouse (*Mus musculus*) arrays. We downloaded close to 7 000 microarrays from the GEO. Of these, 3 715 were human arrays and 3 116 were mouse arrays (representing a total of 11 575 940 possible cross-species microarray matched pairs). We obtained a list of 16 376 human/mouse orthologs from Inparanoid (inparanoid.sbc.su.se).

As was done previously (Le *et al.*, 2010) when performing expression similarity comparisons, we only use the 500 ortholog pairs that vary the most within-species. This reduces the amount of training data required to fit our weight vector. The reason for using this reduced set of orthologs is based on the idea that orthologs which do not vary their levels across experiments in a single organism are less useful in differentiating experiments across species.

The initial positive training set used was our hand-curated set of 138 pairs discussed above. Since we expect the vast majority of random array pairs to represent different experiments, as negative training data we used all pairs not in the positive training set (initially 11 575 802 pairs).

3.1 Cross validation

To determine the ability of our method to recover known similar pairs, and to compare it to prior methods that were based on using only expression values, we performed cross validation on our hand-selected list of positive pairs. In all, 10% of the positive pairs were excluded from the positive set, and then co-training was performed using the remaining training pairs. Pairs were considered to be recovered if they were found in the intersection of the top 5% of the expression and text similarity rankings. Figure 1 presents the performance of our co-training technique compared with the original method from (Le *et al.*, 2010; which is one of the two models in the co-training). As mentioned above, that method only uses expression data and was shown to outperform several other methods that only utilized expression levels (Le *et al.*, 2010). Displayed cross validation scores are an average of 10 runs, each containing a different randomly selected set of excluded pairs. As can be seen, co-training resulted in much higher cross validation accuracy, at 35%, compared with using expression similarity alone, which only

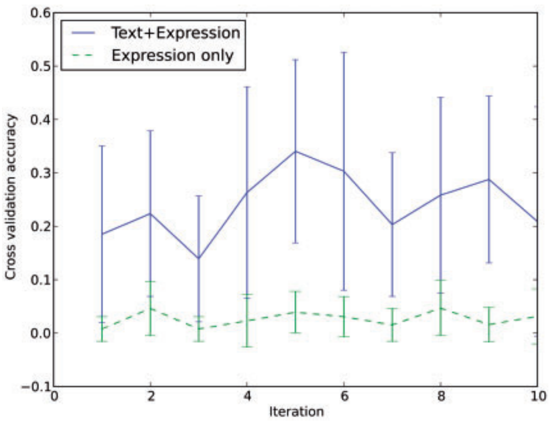

Fig. 1. Cross validation accuracy comparing expression analysis alone to co-training. The co-training method greatly improves between iterations indicating that the new labeled examples contribute to the performance of the combined classifier. Error bars represent standard deviation.

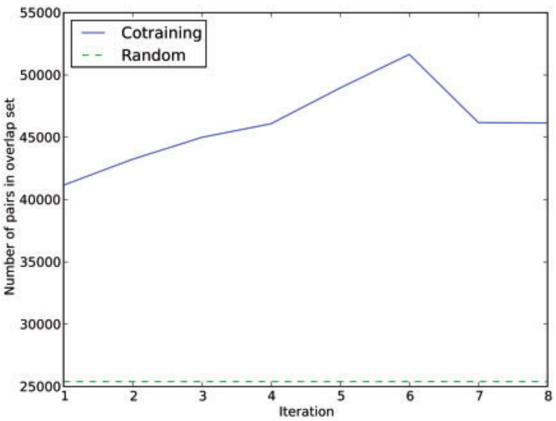

Fig. 2. Overlap between textual matches and expression matches by iteration. Overlap matches (blue) were defined as microarray pairs that were in the top 5% of both most textually similar and most expression similar pairs. Random (green) pairs are pairs chosen from a random 5% of all expression pairs that were also in a random 5% sample of all textual pairs.

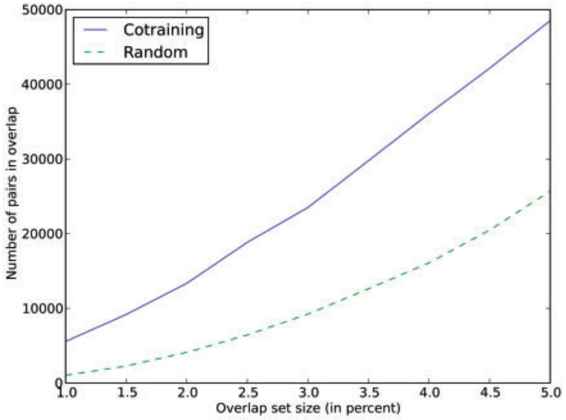

Fig. 3. Overlap between textual matches and expression matches by size of overlap set in iteration 5. Overlap matches (blue) were defined as microarray pairs that were in the top x% of both most textually similar and most expression similar pairs. Random (green) pairs are pairs chosen from a random x% of all expression pairs that were also in a random x% sample of all textual pairs.

leads to 5% accuracy for this dataset. A peak in cross validation performance occurs at the fifth iteration.

Since the initial positive pairs were chosen such that they had high-textual similarity, it is not surprising that co-training (which includes a textual similarity score) outperforms expression similarity alone (which does not) on this dataset. However, it is still noteworthy that co-training increases pair recovery from a baseline of 17% (when using the text data before co-training) to >35%. This indicates that by integrating text and expression our method can improve on the performance of using either one of these datasets on their own. Below we further study the high-scoring co-training matches (both at the dataset and at the array levels) and show how they can be used to derive biological insights about processes and diseases.

Though our cross validation suggests that five iterations is an appropriate training length for our data, this result may be dependent on the number of unlabeled and labeled microarrays and the organisms being compared and so will not generalize for other comparison studies. We have thus also tested an automated method for determining the number of iterations which splits the training data into training, validation and test sets. The first two (together with the unlabeled data), are used to determine the appropriate number of iterations whereas the third is used to evaluate performance. See the Supplementary Website for details.

3.2 Statistical analysis

We performed several analyses of the overlap between textual similarity matches and expression similarity matches. In Figure 2, we show the size of the overlap between the two methods as a function of the iteration of co-training. The overlap consists of all pairs which are determined to be in the top 5% of similarity on both expression and textual metrics. We compare this to the amount of overlap expected if 5% of pairs were randomly drawn from each of the two methods.

It can be seen that there is a substantial enrichment of pairs in the overlap set. By random chance, ∼25 000 pairs should be found in the overlap set. Before any co-training has occurred (iteration 1, where just the initial training data are used) the overlap is 65% larger than random, at 41 000 pairs. At its peak during co-training, the overlap is 51 000 pairs, which is 104% larger than random.

Even though the overlap statistic is largely independent of the cross validation analysis discussed above, similar to the cross validation results, there is a peak in performance at iteration 6. This again suggests that optimal learning occurs after 5–7 iterations.

In Figure 3, we show the size of the overlap set as a function of the percentage of overlap at iteration 5 of the co-training algorithm. That is, at any given point on the *x*-axis, we use that value to determine what top percentage of matches we define as positive.

As can be seen, we consistently obtain a large enrichment in the overlap when compared with random sets of the same size. At the 5% level we have 48 520 pairs from co-training, compared with 25 682 pairs from random.

3.3 Expert evaluation

To further analyze the set of matches determined by our method we selected 100 matched microarray pairs for evaluation. These pairs were randomly selected from the intersection of the top 1% of expression and textual similarities. (All pairs selected were not included in our initial training set.)

Each pair was evaluated on two levels: the correspondence of the experiments (datasets) based on the text abstracts associated with the experiments, and the correspondence of the two individual microarrays that were matched as most similar by our method.

When comparing experiments, we looked at the sort of manipulation performed (i.e. what the experiment was actually testing) as well as what tissue the experiment was performed on. This resulted in the following ranking (higher is better):

(1) Divergent tissue with divergent manipulation;

(2) Divergent tissue with same manipulation;

(3) Homologous tissue with different manipulation;

(4) Homologous tissue with same manipulation or same tissue with different manipulation; and

(5) Same tissue or cell type with (nearly) same manipulation.

On this five point scale, 28% of the pairs were rated 5, 44% 4, 10% 3, 16% 2 and 2% 1. Thus, >70% of the selected matches were scored 4 or 5, indicating that our co-training method was able to successfully match experiments, and conditions, across the two species. We note that random matching leads to 84% of the pairs being scored a 1. Also for the random set, 1% were rated 4, and 1% were rated 5, showing that virtually no high-scoring pairs are expected by chance. See the website for scores for matched and random pairs.

Comparing individual microarrays is more difficult because the information attached to these microarrays in the public databases tends to be limited (often consisting of a couple attributes, such as ‘time: 12 h’ and ‘condition: control’). Thus, we evaluated array pairs on a three-point scale:

(1) Mismatch, divergent condition/time;

(2) Match, similar condition, and homologous tissue (e.g. both of ectodermal origin) or unknown (no tissue info for at least one of the entries); and

(3) Match, same condition and tissue.

Of the 100 pairs, 20% were labeled as 3, 69% as 2 and 11% as 1. The overwhelming number of 2’s is due to the lack of sufficient annotation on the microarrays. However, for pairs that were known, the majority of matches were true positives.

3.4 Cross species analysis of cell cycle genes

To test the usefulness of our new method for determining the function of genes, we used our set of matched array pairs between human and mouse to identify mouse cell cycle genes. As discussed above, when looking at the overlap of the top 5% from each model (expression level and text) we see a significant enrichment in matched pairs; thus we used all pairs in this intersection at iteration 7 (a total of 44 171 microarray pairs). For comparison with prior work, we also selected the top 44 171 pairs using expression level similarity alone.

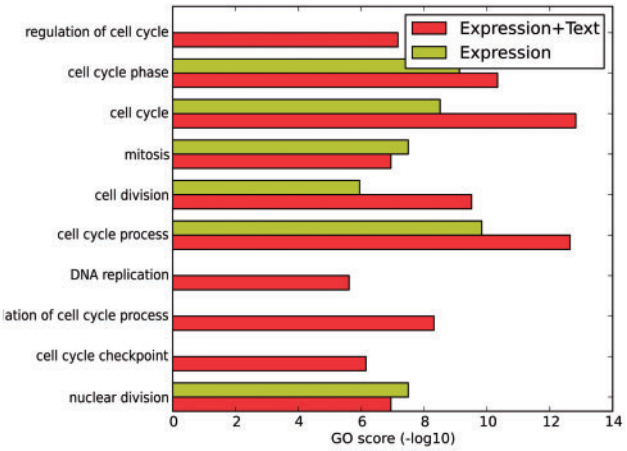

Fig. 4. GO term enrichment of mouse genes with high-expression correlation to cycling genes in putative microarray similar pairs. We show a comparison between the co-training approach and expression similarity alone. All cell cycle related GO terms that are enriched in either of the two sets of genes are included in the figure.

We used a set of 50 human cycling genes identified by (Whitfield *et al*., 2002). For each, we used the set of matched array pairs to select the 10 mouse genes with the greatest Spearman’s correlation resulting in 435 total genes. We used FuncAssociate 2.0 (Berriz *et al*., 2009) to determine GO categories that were strongly enriched in the set of discovered mouse genes. We report all enriched GO terms related to cell cycle activity in Figure 4. These specific GO terms reported were chosen as they were strongly enriched on the initial human set of genes as well. We compare GO enrichment between the co-training method and expression similarity alone. As can be seen in the figure, the co-training method has higher enrichment for most of the cell cycle-related GO terms. For example, for the GO term ‘Cell Cycle Phase’, the co-training method has P-value 4.6E-11 whereas the expression method alone has P-value 7.5E-11; for ‘Cell Cycle Process’ the co-training method has P-value 2.3E-13 whereas the expression method has P-value 1.5E-10. See website for the complete set of enriched GO categories.

3.5 Identifying targets for studying diabetes in mice

After establishing the ability of our method to identify cell cycle mouse genes based on a curated human list, we explored the usage of cross species analysis for studying human diseases. Specifically, we looked at diabetes, a disease affecting 25.8 million people in the USA alone. As a starting point for our cross-species comparison, we selected a set of 19 human genes from KEGG that have mutations known to be associated with type 2 diabetes.

As before, for each of the human genes we selected the top 10 mouse genes with the greatest Spearman’s correlation in the matched array pairs. This resulted in 137 distinct mouse genes (several mouse genes were correlated with multiple human genes, which is expected if the human genes are co-expressed). See the Supplementary Website for a complete list of genes identified by our method. In Table 1 we note some of the most enriched ‘biological process’

Table 1. Top biological process GO terms by *P*-value for mouse genes correlated with human genes that are known to have type 2 diabetes-related mutations

Rank	Category name	Assigned	*P*	*P* adj
1	Developmental process	55	2.19E-17	<0.001
2	Positive regulation of biological process	49	2.4E-14	<0.001
3	Positive regulation of cellular process	46	5.63E-14	<0.001
4	Anatomical structure development	37	1.0E-13	<0.001
5	Anatomical structure morphogenesis	28	6.6E-13	<0.001
12	Positive regulation of metabolic process	29	1.1E-9	<0.001
13	Regulation of metabolic process	51	1.6E-9	<0.001
14	Regulation of primary metabolic process	46	2.3E-9	<0.001
33	Regulation of biosynthetic process	36	3.3E-7	<0.001
96	Positive regulation of immune system process	11	4.0E-6	0.003
96	Regulation of immune system process	13	1.7E-5	0.012
101	Immune system process	14	2.2E-5	0.02

GO terms associated with these mouse genes. As expected, several metabolism and biosynthesis terms were significantly enriched. For example, 'Positive Regulation of Metabolic Process' was enriched with *P*-value 1.1E-9 and 'Regulation of Metabolic Process' was enriched with *P*-value 1.6E-9.

Also notable are several immune-related GO terms including 'Positive Regulation of Immune System Process'. The immune system has been shown to be related to type 2 diabetes. For example, in (Pickup and Crook, 1998) it is suggested that diabetes symptoms (such as the metabolic syndrome that accompanies the disease) is due to a cytokine-mediated reaction and in (Dovio and Angeli, 2001) it is suggested that activity in immune response gene IL-6 is associated with diabetes. Specific immune-related genes, such as TLR4, sCD14 and BPI have known associations to type 2 diabetes (Fernández-Real and Pickup, 2007).

Several of the mouse genes identified by our method either have known association with diabetes or suggest new roles for potential targets of study. For example, IL-18 is known to be produced at elevated levels in patients with type 2 diabetes (Moriwaki *et al.*, 2003) and IL-18 deficiency is also known to lead to insulin resistance in mouse (Netea *et al.*, 2006). Our method identified the mouse gene IL-18r1, which encodes a receptor for IL-18, suggesting a possible direction for overcoming this deficiency by acting directly on the receptor. Another identified gene, CD36, has been associated with diabetic nephropathy, a frequent complication of diabetes (Susztak *et al.*, 2005). Additionally, our subset contained nuclear factor kappa B inhibitor alpha (NFKBIA or IκB), which is known to have polymorphisms associated with type 2 diabetes (Romzova *et al.*, 2006).

Some of the mouse genes were found to be correlated to multiple human genes. For example, intercellular adhesion molecule 1 (ICAM-1) was associated with 7 of the 19 human genes on our list. ICAM-1 has been found to be elevated in diabetic rats (Sugimoto *et al.*, 1997).

4 CONCLUSIONS AND FUTURE WORK

While the availability of genomic sequence data led to several successful computational methods for comparing these datasets across species, relatively little work has been performed to date on mining expression data across species. Given the advantages of expression data (e.g. tissue and condition specificity, and dynamics) developing computational methods for cross species analysis of this data remains an important challenge. Indeed, most drugs are developed and tested using model organisms; our ability to match not just sequence but also the activity of key genes in a specific disease would likely improve the process of drug discovery.

To facilitate such cross species comparisons we developed a novel co-training algorithm that can identify pairs of similar microarrays across species. The value of this approach is two-fold: first, we incorporate textual data into the process of microarray comparison, using a new source of data to better judge array similarity; additionally, we improve the performance of existing expression level similarity models by providing additional, algorithmically selected labeled data.

Testing our method on known similar pairs through cross validation demonstrated that it improves performance when identifying known positive matches. We showed that there is statistically significant overlap between textual similarity and expression level similarity, and that co-training enriches that overlap.

Expert analysis of a subset of our top matches confirmed its accuracy. We next used our matched arrays to identify mouse cell cycle genes as well as mouse genes associated with genes implicated in human diabetes. The list of diabetes-related mouse genes includes several known immune and metabolism genes that play an important role in diabetes as well as novel predictions that can be further tested.

Future work could involve the use of a text analysis technique with a richer parameter space. LSA is only weakly parametrized (by the number of dimensions we retain during dimensionality reduction), and it is likely that we could achieve better performance with an algorithm that can take further advantage of the training examples that are iteratively selected. Additionally, we would like to extend this method to more species pairs, which requires the development of a set of training data and a list of orthologs between the new species pairs.

Funding: NIH and NSF grants NIH [1RO1 GM085022 and NSF DBI-0965316 award to Z.B.J., in part] and NIH T32 training [T32 EB009403 to A.W., predoctoral trainee] as part of the HHMI-NIBIB Interfaces Initiative.

Conflict 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.

Barr,C.S. *et al.* (2003) The utility of the non-human primate model for studying gene by environment interactions in behavioral research. *Genes Brain Behav.*, **2**, 336–340.

Berriz,G.F. *et al.* (2009) Next generation software for functional trend analysis. *Bioinformatics*, **25**, 3043–3044.

Blum,A. and Mitchell,T. (1998) Combining labeled and unlabeled data with co-training. In *Proceedings of COLT 1998*. ACM, New York, NY.

Chapelle,O. *et al*. (2006) *Semi-Supervised Learning*, MIT Press, Cambridge, MA.

Deerwester,S. *et al*. (1990) Indexing by latent semantic analysis, *J. Am. Soc. Inform. Sci.*, **41**, 391–407.

Dovio,A. and Angeli,A. (2001) Cytokines and Type 2 Diabetes Mellitus. *JAMA*, **286**, 2233.

Fernández-Real,J.M. and Pickup,J.C. (2007) Innate immunity, insulin resistance and Type 2 Diabetes. *Trends Endocrinol. Metab.*, **19**, 10–16.

Kaletta,T., and Hengartner,M. (2006) Finding function in novel targets: *C. elegans* as a model organism. *Nat. Rev. Drug Disc.*, **5**, 387–399.

Kuo,D. *et al*. (2010) Evolutionary divergence in the fungal response to fluconazole revealed by soft clustering, *Genome Biol.*, **11**, R77.

Le,H. *et al*. (2010) Cross-species queries of large gene expression databases. *Bioinformatics*, **26**, 2416–2423.

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

Moriwaki,Y. *et al*. (2003) Elevated levels of interleukin-18 and tumor necrosis factor-alpha in serum of patients with type 2 diabetes mellitus: relationship with diabetic nephropathy. *Metabolism*, **52**, 605–608.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Netea,M.G. *et al*. (2006) Deficiency of interleukin-18 in mice leads to hyperphagia, obesity and insulin resistance. *Nat. Med.*, **12**, 650–656.

Pickup,J.C. and Crook,M.A. (1998) Is Type II diabetes mellitus a disease of the innate immune system? *Diabetologia*, **41**, 1241–1248.

Rifkin,S.A. *et al*. (2003) Evolution of gene expression in the Drosophila melanogaster subgroup. *Nat. Genet.*, **33**, 138–144.

Romzova,M. *et al*. (2006) NFκB and its inhibitor IκB in relation to Type 2 Diabetes and its microvascular and atherosclerotic complications. *Human Immun.*, **67**, 706–713.

Rustici,G. *et al*. (2004) Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet.*, **36**, 809–817.

Stark,A. *et al*. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, **450**, 219–232.

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.

Sugimoto,H. *et al*. (1997) Increased expression of intercellular adhesion molecule-1 (ICAM-1) in diabetic rat glomeruli: glomerular hyperfiltration is a potential mechanism of ICAM-1 upregulation. *Diabetes*, **46**, 2075–2081.

Susztak,K. *et al*. (2005) Multiple metabolic hits converge on CD36 as novel mediator of tubular epithelial apoptosis in diabetic nephropathy. *PLoS Med.*, **2**, e45.

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.

Zinman,G. *et al*. (2011) Large scale comparison of innate responses to viral and bacterial pathogens in mouse and macaque. *PLoS ONE*, **6**, 7:e22401.