Gene expression

Cross species analysis of microarray expression data

Yong Lu1,†, Peter Huggins2 and Ziv Bar-Joseph1,2,*

1School of Computer Science and 2Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Received on January 21, 2009; revised on April 3, 2009; accepted on April 5, 2009

Advance Access publication April 8, 2009

Associate editor: Jonathan Wren

ABSTRACT

Motivation: Many biological systems operate in a similar manner across a large number of species or conditions. Cross-species analysis of sequence and interaction data is often applied to determine the function of new genes. In contrast to these static measurements, microarrays measure the dynamic, condition-specific response of complex biological systems. These recent exponential growth in microarray expression datasets allows researchers to combine expression experiments from multiple species to identify genes that are not only conserved in sequence but also operated in a similar way in the different species studied.

Results: In this review we discuss the computational and technical challenges associated with these studies, the approaches that have been developed to address these challenges and the advantages of cross-species analysis of microarray data. We show how successful application of these methods lead to insights that cannot be obtained when analyzing data from a single species. We also highlight current open problems and discuss possible ways to address them.

Contact: zvbj@cs.cmu.edu

1 INTRODUCTION

Several biological systems operate in a similar way in diverse set of organisms and many of the genes that participate in these systems are conserved across these organisms (Wilkins, 2001). Sequence similarity is one of the major sources of data for determining the function of new genes, e.g. by using BLAST (Altschul et al., 1990). Comparative genomics, a term used to describe large-scale comparisons of complete genomes, has aided in the correct identification of the genes and control regions in each of the sequences being compared (Kellis et al., 2003) and sequence conservation analysis led to the identification of hundreds of new miRNAs (Stark et al., 2003). Other efforts focused on cross-species analysis of interaction data, identifying core interaction modules as well as differences between regulatory programmes in closely related species (Odum et al., 2007; Sharan et al., 2005).

While useful, sequence conservation analysis and network comparisons only tell part of the story. Sequence data does not change during the time of operation of the biological systems. While interactions may change between conditions and time, almost all current interaction data is static focusing on one time point and one condition (Harrison et al., 2004; Krogan et al., 2006). Thus, using only these datasets it is often hard to tell which genes participate in various biological processes. In addition, in some cases large changes in sequence and interactions may only have a minor affect on function whereas in other cases small changes in sequence between two genes may result in large changes in structure leading to divergent function for the genes (Alexander et al., 2007).

To address these issues researchers use microarrays to measure the dynamic, condition-specific response, of complex biological systems involving hundreds of interacting proteins. Examples include the cell cycle (Spellman et al., 1998), immune and other stress responses (Nau et al., 2002), circadian rhythm (Correa et al., 2003) and developmental processes (Arbeitman et al., 2002). These processes are shared between multiple, and in some cases distant, species. By combining and comparing these experiments across species we can identify a ‘core’ set of genes. These genes are conserved in both sequence and expression between multiple species and are thus key components of the biological response or system being studied.

2 COMPUTATIONAL CHALLENGES IN THE ANALYSIS OF DATA ACROSS SPECIES

Many of the successful applications of cross-species analysis to sequence and interaction data utilized powerful computational techniques. Combinatorial and probabilistic methods were developed to align sequence data (Altschul et al., 1990; Durbin et al., 1998) and graph-based algorithms are used to carry out whole genome alignments (Kamvysselis et al., 2003). More recently, researchers developed sophisticated computational methods for cross-species comparisons of interaction networks (Sharan et al., 2005).

When comparing microarray datasets across species, researchers face many of the same challenges that arise when comparing other high throughput datasets including the search issues related to the large datasets and the need to handle homology assignments between species. In addition, a good experimental design that takes into account the fact that experiments are to be compared across species is crucial for the success of such studies. However, microarrays also raise several new challenges. Microarray data is often noisy. The agreement between experiments measuring similar processes in different labs, even within the same species, is sometimes very small (Oliva et al., 2005). Another challenge results from the differences in conditions and dynamics. Unlike sequence or interaction data which are often represented by a small number of letters (DNA)
We next discuss applications that use the same microarray to
co-regulation information in one species to improve expression
species meta-analysis can be used to leverage annotation and
has also been performed on data from multiple species. Cross-
recently, with the increase of cross-species data such meta-analysis
is to search for similar genes in other species. Similarly with the growth in microarray datasets applications that can combine data across species may lead to
new insights into biological systems that are activated in multiple species.

Many single species studies combine data from multiple microarray
experiments (Ramaswamy et al., 2001; Spellman et al., 1998). Recently, with the increase of cross-species data such meta-analysis
has also been performed on data from multiple species. Cross-
species meta-analysis can be used to leverage annotation and
coregulation information in one species to improve expression
analysis in a less-studied species (Rustici et al., 2004), find common
expression patterns in multiple species to reveal core gene functions
(Stuart et al., 2003) and elucidate the evolution of gene expression
and coregulation (Bergmann et al., 2004). Cross-species meta-
analysis can be roughly divided into two types: coexpression
meta-analysis and expression meta-analysis. These two approaches
study inherently different questions. Coexpression meta-analysis
asks whether genes coexpressed in one species are also coexpressed
in another species. Expression meta-analysis directly analyzes
the similarity between expression profiles of homologous genes
in different species. Thus, while expression analysis identifies
homologous genes that respond in the same way to specific stimuli
in multiple species, coexpression analysis can result in genes with very
different response patterns in each of the species. On the contrary,
coexpression analysis allows the use of different conditions for the
different species studied whereas expression analysis requires the
use of the same conditions in all species.

3 Coexpression meta-analysis

The first method that has been applied to study cross-species
microarray experiments is coexpression meta-analysis. Rather than
compare expression data directly between species, evidence for gene
coeXpression is derived separately in each individual species, and
then combined to infer gene modules.

One advantage of coexpression meta-analysis is that microarray
experiments for different species can be combined even under
different experimental conditions. In Stuart et al. (2003), the
concept of a metagene is introduced (Fig. 3 Top). A metagene is
a set of strictly orthologous genes among multiple species. Two
metagenes are declared to be coexpressed if their constituent genes
are significantly coexpressed in each of the species; then modules
of coexpressed metagenes can be identified, which can be regarded
as ‘core’ modules of genes. In Bergmann et al. (2004), a ‘signature
algorithm’ is presented which maps an annotated gene module A in
one species to its set of homologous genes B in a second species, and
then coexpressed genes in B are extended to a gene module which
Fig. 2. Strategies for combining and comparing microarray datasets from multiple species. Left: expression meta-analysis. Samples from each species are hybridized to different arrays and each array is independently analyzed to identify differentially expressed genes. Lists of such genes are later compared to identify the overlap. Middle: using the same array for all species. Samples from all species are hybridized to the same array and all arrays are analyzed using the same method. The list of differentially expressed probes can then be compared. Note that this method can only be used to compare closely related species. Right: concurrent analysis of expression data. Samples from each species are hybridized to separate arrays but are analyzed together so that homologs can be used to improve the assignment of genes.

includes any additional genes which are coexpressed with B (Fig. 3 Bottom). Both methods identified significant co-expression modules but also revealed cases of divergent coexpression between species. Closely related to the signature algorithm, the software OSCAR (Lu et al., 2007b) provides tools for clustering and transferring cluster information from one species to another. A website that can facilitate coexpression meta-analysis is the yeast Microarray Global Viewer (Lelandais et al., 2004) which allows users to view expression values for the genes of budding and fission yeast under different experimental conditions. By studying species ranging from bacteria to human, researchers have also identified dynamic expression patterns that are highly conserved across species (Ueda et al., 2004).

3.2 Expression meta-analysis
The second method, expression meta-analysis, directly compares the expression of orthologous genes under similar conditions or in the same tissue.

Most expression meta-analysis efforts focused on comparing lists of differentially expressed genes (DEGs) derived from published papers (Fortunel et al., 2003; Han and Hickey, 2005). For example, lists of DEGs from 22 dietary restriction studies spanning six organisms were compared by Han and Hickey (2005). They found no agreement in DEG lists for different species, and very little agreement even within species. Out of 15 studies on dietary restriction in mouse, only one DEG was found by four studies, and no DEGs were identified by five or more studies. Databases such as LOLA, L2L and MSigDB (Cahan et al., 2005; Newman and Weiner 2005; Subramanian et al., 2005) have been created to store DEG lists from microarray studies, to facilitate list-based meta-analysis.

In this context it is important to note that directly comparing lists of ‘significant’ genes (even when using the same P-value cutoff for all papers) can exaggerate the disagreement between studies, because genes identified as significant in one study might only be weakly significant in other studies. This is especially true if the different studies used different methods to determine a P-value, which is often the case. More sophisticated methods for combining and comparing microarray data from related studies in a single species are reviewed by Hong and Breitling (2008). Three different types of methods are discussed: raw aggregation, rank-based and P-value based. These methods can be adapted to cross-species data, provided a one-to-one orthology relationship is known.

Raw aggregation methods combine the expression data for a gene from each microarray experiment, and then evaluate the
of 0.392) (Liao and Zhang, 2006a). Analysis of exon arrays from six overlap between orthologs (Spearman's rank correlation coefficient correlation rather than pre-selecting sets of genes found a significant However, later analysis of this data focusing on expression little overlap between the tissue expression of orthologous genes.

of species 2 can then be extended (\(b\)) if their constituent gene pairs are sufficiently coexpressed. Bottom: biclustering methods (Bergman et al.; 2004; Lu et al., 2007b). Given a set of coexpressed genes in species 1 (\(a\)), orthologous genes are identified in species 2 which are coexpressed in species 2 (\(b\)). The set of genes in species 2 can then be extended (\(e\)) to include additional coexpressed genes.

significance of the combined expression data. Error variances can be included to account for global systematic differences between different experiments, and also for noise within each experiment. Choi et al. (2003) developed such an aggregation method, termed a \(t\)-based approach, and applied it to various cancer datasets. They demonstrated that slight but consistent expression changes could be identified for some genes, which no one microarray study alone could identify as significant. Rank-based methods sort the genes in each study according to the significance of differential expression and then aggregate the rankings of each gene across studies, which allow to overcome differences in \(P\)-value comparisons. Permutation tests can then be used to identify genes whose aggregated rankings are significant. Finally, \(P\)-value aggregation methods combine the \(P\)-values for each gene’s expression obtained from different microarray experiments, to obtain an aggregate \(P\)-value. \(P\)-values for a gene can be aggregated by various methods, e.g. by taking the minimum or the product of the \(P\)-values.

Recent studies point out that using DEGs and strict cutoff methods may lead to underestimating expression conservation. Initial analysis of a large-scale comparison of gene expression in over 50 mouse and human tissues (Su et al., 2004) identified little overlap between the tissue expression of orthologous genes. However, later analysis of this data focusing on expression correlation rather than pre-selecting sets of genes found a significant overlap between orthologs (Spearman’s rank correlation coefficient of 0.392) (Liao and Zhang, 2006a). Analysis of exon arrays from six human and mouse tissues resulted in even higher correlation (0.69) (Xing et al., 2007).

3.3 Indirect comparisons

Direct cross-species comparison of DEGs or correlation is more challenging in distant species. Instead, Kumar et al. (2005) computed enriched GO terms for the DEGs in each species, and then looked for enriched terms that are conserved across species. This method was applied to microarray data obtained from mouse, rat and dog treated with LPS (a bacterial cell wall material). They found that although very few orthologous genes could be identified as differentially expressed, GO terms such as Xenobiotic Metabolism were enriched in all three species.

The method of Kumar et al. treats each gene dichotomically, as either differentially expressed or not. Subramanian et al. (2005) proposed a more sophisticated approach which they called gene set enrichment analysis (GSEA). The set of genes G in a microarray are first sorted according to a differential expression score. Then, for a predefined subset of genes of interest (e.g. a GO category), an enrichment score is calculated which essentially measures how disproportionately high the scores of the top-ranked genes in the subset are among the top-ranked genes in G.

In GSEA the sets of genes can be defined in different ways, e.g. by GO categories, results of previous microarray studies, or pathway data. Tools such as DAVID, GenMAPP and WholePathwayScope (Huang et al., 2009; Salomonis et al., 2007; Yi et al., 2006) allow the user to compare a list of DEGs to predefined sets of genes, to look for significantly enriched annotations. Enriched annotations can then be compared across species. Identifying enriched pathways from lists of DEGs has been shown to improve reproducibility and comparability of microarray studies of prostate cancer (Manoli et al., 2006), and may also help facilitate comparisons of microarray studies across species.

Liu et al. (2007a) extended GSEA to gene network enrichment analysis (GNEA). In GNEA, a highly scoring subnetwork is found within a given protein–protein interaction network. Then predefined gene sets (e.g. GO categories) are tested for enrichment within these subnetworks. Applying GNEA to a combination of type 2 diabetes datasets from human and mouse, Liu et al. implicated gene sets involved in insulin signalling and nuclear receptors as differentially expressed in diabetes. Previous analysis of type 2 diabetes data failed to identify any common gene sets, due to the complex and multifactorial nature of the disease.

4 USING THE SAME ARRAY FOR MULTIPLE SPECIES

The methods discussed above compare data from species–specific arrays. A problem with these methods arises from the different probe sets used by each array which may have different hybridization properties. This probe-dependent expression measurement, often termed probe effect, will bias the estimation of gene-expression levels (Irizarry et al., 2005) leading to over-estimation of the disagreement between species. For effective comparison of gene expression across species, it is desirable to control for platform-related variations. One way to circumvent this issue is to use the same microarray to study different species.

There are two primary strategies for using a single array type when studying multiple species: using an array constructed for a single species (Bigger et al., 2001; Rifkin et al., 2003) and constructing a customized array containing probes for every species studied.
microarrays designed for *Drosophila melanogaster* for a closely related species. For example, Nuzhdin able to hybridize with its ortholog. species. As a result, probes designed for a gene in one species are able to hybridize with its ortholog.

There are a number of advantages for using single species microarrays. It is feasible even if genomic data is only available for a closely related species. For example, Nuzhdin et al. (2004) use microarrays designed for *Drosophila melanogaster* to study gene expression in *D. simulans* and *D. melanogaster*. They found that 48% of the genes represented on the array are expressed in both species, and 7% of the commonly expressed genes have significantly different transcript levels (\( P < 0.0001 \)).

Another application is in preclinical cancer drug screening, where animal models with transplanted tumours are often used to study the progression of tumours, identify therapeutic targets and test treatment response. Single species microarrays can be used to validate the animal model by comparing gene expression in both the animal model and the primary tumour in human tissues (Whiteford et al., 2007).

There are several issues to consider when using single species arrays to study multiple species. The first is to decide which species the microarrays should be based on. Because the probes are designed for one species and used to detect genes in another species, in some cases the sequence of a probe does not match that of the target gene, which may result in weaker hybridization. This effect, which is due to sequence mismatch, complicates the estimation of gene expression levels (Gilad et al., 2006; Sartor et al., 2006). To alleviate the sequence mismatch effect, it is desirable to select a species that is as close as possible to the target species.

When studying more than two species using the same single species microarrays, the sequence mismatch effect varies for each species due to different evolutionary distances between them. Gilad et al. (2005) compare the sequence mismatch effect on four species (human, chimpanzee, orangutan and rhesus macaque), and show that, as expected, sequence mismatch effect becomes more severe as sequence divergence increases. The mismatch effect is significant in 63% (577 out of 912) genes when sequence divergence between species is 1% (human and chimpanzee), rising to 89% (759 out of 851) genes when sequence divergence is 5% (human and rhesus macaque). The variations make it more difficult to compare gene expression directly.

A possible way to avoid this problem is by masking probes with a sequence mismatch between orthologous genes in the species studied (Khaitovich et al., 2005; Kirst et al., 2006). The drawback is that when comparing more than two species, a large amount of the probes will have to be masked, severely limiting the number of genes that can be measured in the study (Oshlack et al., 2007).

The main computational problem in the application of single species arrays is to develop normalization methods that can account for the effect of sequence mismatch and of cross-hybridization. Both effects are gene dependent because of the various divergence rates among gene families. It remains an open question to develop methods that can correctly normalize the probe measurements to account for these effects.

### 4.2 Multi-species array

A second approach is to use a microarray containing probes for orthologous genes from every species under study. In a multi-species array, samples from two species are competitively hybridized to the probes. The expression level of a gene is estimated by averaging the log-ratios reported by probes from the same species as well as those from the other species. Oshlack et al. (2007) shows that multi-species microarrays can alleviate the problem of sequence mismatch effects, when compared with species-specific arrays. Their idea is that, under the assumption that sequence mismatch effects are symmetric between the two species, taking the average log-ratio will cancel out the bias caused by cross-species hybridizing. These arrays can be used to study expression of evolutionarily conserved genes. For example, Veltchev et al. (2006) compared the gene expression in the oocyte of bovine, mouse and Xenopus laevis using multi-species microarrays representing 3456 genes, and concluded that 268 (8%) of the genes are preferentially expressed in the oocyte in all three species.

Because probes from multiple species are used, their hybridizing properties will vary. If the probe effect and the sequence mismatch effect are additive then averaging log-ratios from probe sets across species is not enough to completely remove the noise due to cross-species hybridization, especially when the probe effect is significant and the distance between species are large. In fact, it is estimated that probe effect is significant for 59% of genes represented in a study on human and rhesus macaque (Oshlack et al., 2007). As mentioned above, a key problem is to develop computational methods that can adjust for the effect of both the sequence mismatch and probe properties.

Multi-species arrays are only suitable for genes with an ortholog in every species under study. Consequently, the genes one can study are limited by the distance between species. For example, Liao and Zhang (2006b) determined that only 10 607 ortholog pairs between human and mouse are suitable for a multi-species array study, which covers less than half of human proteins and less than 40% of the known mouse genes. For more distantly related species, the coverage will be even lower.

### 5 COMBINED ANALYSIS OF MICROARRAY DATA FROM MULTIPLE SPECIES

A few recent methods combine ideas from both previous approaches. Similar to the first set of methods they use a species-specific array for each species. Similar to the second set of methods the analysis of all data is done concurrently. These methods have primarily been applied to study the cell cycle, though they are general and can be applied to other biological systems.

Early applications of microarrays to study the cell cycle focused on the identification of cycling genes in different species including budding yeast (Spellman et al., 1998), human (Whitfield et al., 2002), plants (Menges et al., 2002) and bacteria (Laub et al., 2000). In 2004, researchers profiled *S. pombe* (fission yeast) and used an expression meta-analysis method (discussed above) to compare the resulting cell cycle patterns to those from *S. cerevisiae* (budding yeast). The results were surprising. Rustici et al. (2004) concluded...
that less than 10% of cycling genes in either species had cycling homologs in the other species. However, comparisons using data generated in other labs led to different conclusions. For example, Oliva et al. (2005), again using expression meta-analysis, identified more than 30% overlap in the top list of cycling genes between the two yeasts.

Following these initial comparisons, Ota et al. (2004) looked at conservation between yeast, human and Arabidopsis concluding that more than 10% of genes are cycling and conserved between these very distant species. Dyczkowski and Vingron looked at cell cycle agreement between the two yeasts and humans (Dyczkowski and Vingron, 2005) and determined that only a small fraction of genes (ranging from 2–8% depending on the species) are conserved in all three species. Jensen et al. (2006) compared all four species (the two yeasts, plants and human cell cycle) identifying less than 1% of orthologous groups as cycling in all four species. Still, the combined analysis of expression data from all species did lead to another important finding. Jensen et al. determined that while expression may not be conserved, the set of complexes involved is. In different species, different units of the complex cycle lead to similar utilization during the cell division process.

The above studies differed in the method used to determine whether a gene was cycling or not but they all used similar expression meta-analysis techniques. Data from each species was analyzed independently and results were compared using curated database assignments. While these methods allow for unbiased analysis of the results they also suffer from a number of drawbacks that can lead to lower overlap between the different species as discussed above. Another challenge is the need to define a one-to-one orthology relationship to carry out such an analysis. The binary assignment (ortholog or not) in databases cannot account for more complex similarity measures which are often represented using a more continuous value (e.g. Blast e-value).

Alter et al. (2003) was among the first to concurrently analyze cell cycle expression data from multiple species. Using generalized singular value decomposition (GSVD) for comparing human and budding yeast cell cycle datasets, the authors were able to identify common and unique response patterns and to recover more accurate cell cycle expression profiles for some of the genes based on the information from the other species. However, this method did not attempt to identify specific genes that are conserved in both sequence and expression profile. Lu et al. (2006) used Markov random fields (MRF), an undirected graphical model, to concurrently analyze cell cycle data from human and budding yeast. In these models genes from all species (and their expression patterns) are represented as nodes and sequence similarity is encoded by edges (Fig. 4). Information is propagated along the edges allowing genes to influence the assignments of their homologs. Thus, instead of using a strict threshold, they use a flexible threshold which can be adjusted based on homology information.

The method affects genes with borderline scores, elevating those more continuous value (e.g. Blast complex similarity measures which are often represented using a assignment (ortholog or not) in databases cannot account for more discussed above. Another challenge is the need to define a one-to-one orthology relationship to carry out such an analysis. The binary assignment (ortholog or not) in databases cannot account for more complex similarity measures which are often represented using a more continuous value (e.g. Blast e-value).

Lu et al. (2006) have reanalyzed the expression data from the four species mentioned above and concluded that the agreement between the two yeasts was 20–25% and that 5–8% of cycling genes in all species have cycling homologs in all other species.

While the combined analysis method led to improved overlap, it may have overstated the set of conserved cycling genes because it utilizes homology information to influence cyclic assignments (Jensen et al., 2008). To independently test the importance of the combined analysis, Lu et al. showed that 45% of yeast genes determined to have cycling homologs in all species were essential. In contrast only 18% of all yeast genes are essential and only 16% of cycling yeast genes are essential. Similar results can be shown for human genes using RNAi analysis. Analysis of DNA motif data further supported these findings indicating that at least part of the conserved expression results from conserved regulation (Jensen et al., 2008). Thus, by combining sequence and expression data in a unified analysis across species we can identify genes that represent the core functional units of the cell cycle programme.

6 DISCUSSION

The exponential growth in microarray datasets over the last decade opens the door for large-scale, cross-species comparison studies. Similar analysis of sequence and interaction data has led to many important findings and the algorithms and computational tools developed for these comparisons are routinely used.

Analysis of cross-species microarray data is challenging. Direct comparison of these experiments requires that they would be carried out in a very similar manner in all species and temporal differences between the species need to be accounted for prior to the actual comparisons. Still, many such studies were carried out in closely related species. These studies identified common and unique expression patterns used during development and in specific tissue types. They have uncovered conserved functional categories and interactions networks that are commonly activated in the different species. Table 1 summarizes the methods that have been suggested for analyzing such experiments.

A few biological systems, including the cell cycle, immune response and a number of diseases were studied in more distant species and these are primed for large-scale comparison studies. So far, the cell cycle has received the most attention, perhaps because of the initial surprising findings of very low conservation between two yeast species. Researchers have also shown that some aspects of stress response, including the repression of ribosomal genes, are significantly conserved across a large number of yeast species (Lelandais et al., 2008; Tirosch et al., 2006). More recently researchers have also looked at a number of diseases, including diabetes (Liu et al., 2007b), and while only few genes were determined to be commonly expressed between species, a number of subnetworks were identified as common responses indicating that expression may be conserved in different units of the same network.

Initial analysis of cross-species microarray data used different conditions in different species to study global conservation of coexpression within species (Stuart et al., 2003). If such an approach can be extended to allow for the identification of similarly expressed genes between species (while still using conditions that are not identical) it could open the door to many new applications. Consider a BLAST like search through Genbank and GEO for genes from multiple species coexpressed with a query gene. Such genes will have both sequence and functional similarity to the gene of interest and can lead to better assumptions regarding its function. The computational challenge would be to determine which experiments in one species correspond to experiments in another species so that they could be queried in such searchers. One approach would be to manually curate these large datasets...
Concurrent analysis of microarray data from multiple species (Lu et al., 2007a). Sequence data (left) and expression data (bottom right) are used to construct a graph in which gene expression is represented by nodes and edges represent homology (top right). Information is propagated along edges in this graph allowing genes to influence the assignment of their homologs. This can be used to elevate a borderline gene if many of its homologs are strongly expressed overcoming problems related to noise and treatment differences. The set of edges can be derived from a curated database or from sequence analysis methods including BLAST.

Table 1. Comparison of major features of methods developed for cross species expression analysis

<table>
<thead>
<tr>
<th>Coexpression meta-analysis (metagene, signature algorithm)</th>
<th>Expression meta-analysis (raw aggregation, P-value)</th>
<th>Indirect (GSEA)</th>
<th>Single species arrays</th>
<th>Multi-species arrays</th>
<th>Combined analysis (MRF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single platform for all experiments</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Customized array</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Similar experimental conditions</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Directly compares expression profiles</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Can be applied to distant species</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Requires orthology information</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Separate P-value cut-off for each species</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>


to identify closely related conditions. A more promising direction would be to automatically extract such similarity knowledge based on a small set of known orthologs. If these queries prove successful they can change the way researchers study their genes of interest.

**Conflict of Interest**: none declared.

**REFERENCES**


