Systematic Measurement of Missmatch Effect for Designing Inter-Species Microarray

Mutsumi Fukuzaki
Dept. of Computer Science,
Ochanomizu University
Tokyo, Japan
fukuzaki@sel.is.ocha.ac.jp

Masa-aki Yoshida, Atsushi Ogura
Ochadai Academic Production,
Ochanomizu University
Tokyo, Japan
yoshida.masaaki@ocha.ac.jp
aogu@whelix.info

Jun Sese
Department of Computer Science,
Tokyo Institute of Technology
Tokyo, Japan
sesejun@cs.titech.ac.jp

***Abstract*—After the success of comparative genomics studies, comparative transcriptomics have gained in importance. A number of methodological drawbacks, however, prevent us from fully understanding expression evolution: target species are limited to very closely related species, and the number of target genes is small. In this study, we propose a new microarray to measure multi-species gene expression levels, in which orthologous genes are measured with a single probe. We checked the effects of probe sequence mutations on expression levels using systematically mutated probes, and found that mutations on the 3'-end of probes have very little effect on expression values whereas those on the 5'-end decreased the observed expression levels. According to the results, we generated an array for humans, rats and mice that covered 6683 genes. Both the species-range and the number of genes are larger than those of previous arrays.**

***Index Terms*—Comparative Transcriptomics; Microarray; Gene Expression; Evolution**

I. Background

Comparative genomics have been used to gain a better understanding of evolutionary processes. Recent whole genome comparison studies of closely related species have revealed few major differences in the distribution of their gene functions [1]. The results have moved multiple-species gene expression studies forward to uncover the evolution of gene expression and have helped to identify genes that evolved under selective pressures [2], [3] as well as similarities in developmental patterns over multiple-species [4]–[6]. Recent studies have used microarrays to compare gene expression between closely related species on a large scale [5]–[7]. However, because microarrays are currently available for only a limited number of species, most studies have assayed multiple species using arrays designed on the basis of sequence data from only a single species. For example, *Drosophila melanogaster* arrays have been used to directly compare expression levels in *D. melanogaster* and *D. simulans* [8], and *Arabidopsis thaliana* arrays have been used to compare expressions between species closely related to *A. thaliana* [9], [10].

Several papers have reported that sequence mismatches create a problem for gene-expression measurement in different species when using single-species arrays [11]–[13]. As the probes on most arrays are designed using sequence data from only one species, they can differ by many base pairs from the target cDNA derived from other species. Ranz *et al*. [8] estimated the effect of sequence divergence between *D. melanogaster* and *D. simulans* by hybridizing genomic DNA from both species to the array. Gilad *et al*. [13] developed cDNA arrays containing probes from multiple species to compare humans, chimpanzees, orangutans and rhesus monkeys, and evaluated the array. However, the cDNA arrays used in these experiments were more expensive than recent commercially available microarrays, such as Agilent, NimbleGen (Roche) or Affymetrix arrays, because the studies used a cDNA-spotting machine. Utilization of commercial arrays will allow the problem of cost to be avoided.

Recent expression studies have utilized new generation sequencers known as RNA-Seq [14]. In comparison with new generation sequencers, microarray experimental protocols are more established, and analysis procedure is simpler.

In this paper, we introduce a microarray that can be used to observe gene expression over multiple species using the Agilent custom microarray. We first show that the presence of mutations close to the 3'-end of probes has very little effect on expression levels, while the presence of a single mutation close to the 5'-end of probes has a large effect on observed expression levels. We then introduce a microarray design method that identifies probes hybridizing target genes but not hybridizing non-target genes. As an initial example, we applied the method to measure human, rat and mouse expression levels with one microarray covering 6,683 genes (ortholog groups). The number of target genes and the range of species are both wider than the array presented by Gilad *et al*. [13].

II. Materials and Methods

A. Artificial Mutation

We used an Agilent custom microarray with a probe length of 60bp. To check the effects of artificial mutations, we designed 56 patterns of mutations, and 24 patterns of insertions and deletions (Details are described in Section II-B). On the microarray, we used 7,987 probes from the Agilent Human GE 4x44k v2 Microarray whose associated genes were expressed in the BioGPS database [15], as our purpose was to check for changes in gene expressions, thereby requiring genes

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

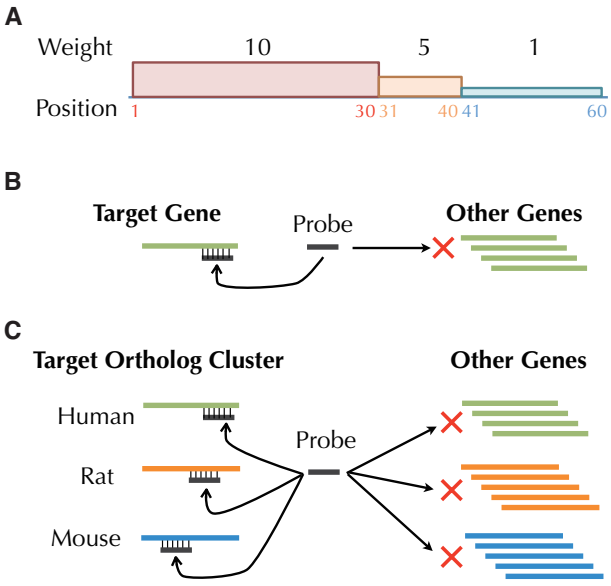

Fig. 1. Overview of probe design. (A) Weighted distance. To assess the hybridization of probes to a target sequence, we defined the weighting on each position. We set weighting on positions 1 to 30, 31 to 40 and 41 to 60 from the 5'-end in probes as 10, 5 and 1, respectively. This weighting and hybridization threshold means that a probe having a single mutation between position 1 to 30 does not hybridize to the target gene because its weighted distance from the target gene is 10, whereas a probe having five mutations between position 41 to 60 can hybridize to the target gene because the distance is 5. This result corresponds to those of the artificially mutated probe experiment in Fig. 3. (B) Probe design for a single-species microarray. (C) Probe design for an inter-species microarray. Each probe has to hybridize to all genes in a target ortholog cluster.

expressed in human cells. In these experiments, we checked the differences in expression level in at least 73 probes for each mutation type.

To check the effect of mutations on expression levels, we generated two different microarrays. One contained only probes for which the sequences are identical to human genome, while the other contained the mutated probes described in the previous section. The non-mutated and mutated arrays contained 7,987 and 8,268 probes, respectively.

We normalized them according to the Agilent user manual with 219 probes that were included in both arrays. In the microarray experiment, we used qPCR Human Reference Total RNA obtained from Clonetech. We repeated 6 experiments for both non-mutated probes and mutated probes to obtain stable values.

B. Mutation Positions

We randomly assigned the patterns to probes and changed probe sequences according to the patterns. The sequences are artificially mutated. When a nucleotide was mutated, A, C, T and G were changed to T, G, A and C, respectively, as these changes do not affect the TM value of the probe sequence.

C. Weighted distance between sequences

We used weighted distance to investigate whether probes hybridize to target sequences or not. The weighted distance between 60-nucleotide probes $\vec{p} = p_1 p_2 \cdots p_{60}$ and n-nucleotide gene sequences $\vec{g} = g_1 g_2 \cdots g_n$ was calculated as shown below. To check the hybridization of two sequences, we considered the minimum distance between the sequences as that of the closest sequences that will hybridize to each other when the probe and gene hybridize. We first determined the weight of each position on the probe as follows:

$$w(i) = \begin{cases} 10 & (i \leq 30) \\ 5 & (30 < i \leq 40) \\ 1 & (40 < i) \end{cases}$$

Fig. 1(A) illustrates the weightings. With this weighting, we calculated the weighted distance $D(\vec{p}, \vec{g})$ between two sequences $\vec{p}$ and $\vec{g}$ as follows:

$$D(\vec{p}, \vec{g}) = \min\{d \mid d \text{ in } d(p_{60}, g_j) \text{ for } 60 \leq j \leq n\}, \text{where}$$

$$d(p_i, g_j) = \min \begin{cases} 0 & (i \leq 0) \\ \sum_{k=1}^{i} w(k) & (j \leq 0) \\ d(p_{i-1}, g_j) + w(i) & \\ d(p_i, g_{j-1}) + w(i) & \\ d(p_{i-1}, g_{j-1}) + w(i) & (p_i \neq g_j) \\ d(p_{i-1}, g_{j-1}) & (p_i = g_j) \end{cases}$$

This distance can be efficiently calculated using a dynamic programming algorithm.

D. Inter-species microarray design

We used the Homologene database [16] to obtain groups of orthologous genes between humans, rats and mice. The Homologene database was downloaded on April 5th, 2010 and we selected groups containing a human gene, a rat gene and a mouse gene. The number of groups used was 14,217.

From the 14,217 groups in the Homologene database, we developed a new program to find sequences that can be used for microarray probes. We first selected probes that could measure gene expression in all three species. The procedure generated 6,683 probes. We then designed probes to measure gene expression in humans and mice but not to hybridize to rat genes, and generated 1,400 probes. Similarly we designed probes for rats and mice and for humans and mice. Finally, we designed probes for each gene for which we had not previously generated a probe. The numbers of probes designed are given in Fig. 2. We generated 20,914 probes for 13,498 (94.9%) out of 14,217 ortholog groups (genes).

III. Results and Discussion

A. Probes having artificial mutations and changes in their expression levels

Microarray probe design is similar to PCR primer design in that both probe and primers hybridize to complementary sequences. The primer design problem has already been studied both practically and theoretically [17], [18].

The theoretical problem of primer design for a gene was defined as the need to find a sequence whose complementary sequence exists in the target gene but not in any non-target genes. To date, probes have been selected from completely complementary sequences of each target gene. The extension of these criteria theoretically allows us to design microarray probes to measure multiple-species; that is, to select a

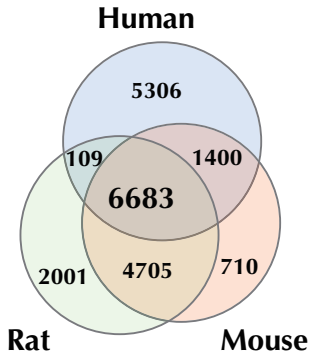

Fig. 2. A Venn diagram showing the number of probes for each of the three species.

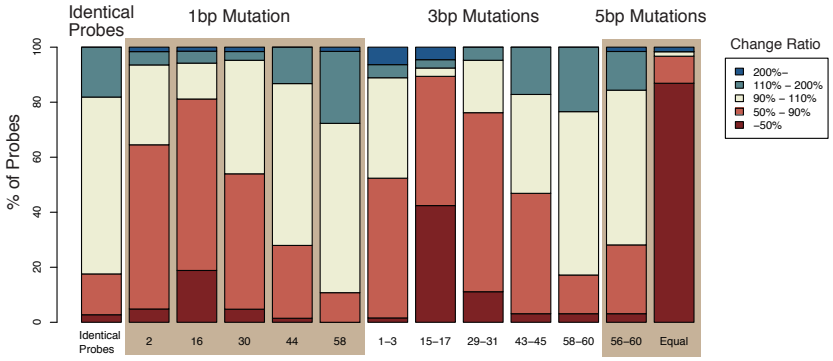

Fig. 3. Effects of artificial mutations on probes. We visualized the changes in expression level between two independent microarrays: a non-mutation original array and an array having artificial mutations. Bars indicate mismatch types and labels below bars indicate the mutation positions. For example, the bar labeled "1-3" shows the differences in expression of probes having mutations at the 1st, 2nd and 3rd nucleotide from the 5'-end region. Each bar represents the changes in expression level. We categorized the change into 5 categories and the heights of the color bar indicate the ratios of the probes in the categories. The large cream region (90%-110%) represents probes for which the expression level did not change after insertion of a mutation(s). Large red region represents probes for which the expression level was decreased after insertion of a mutation(s) in comparison with those of the original probes. We measured more than 73 probes for each mutation variation. All experimental procedures are described in Sections II-A and II-B.

sequence existing in all target orthologous genes but not in other genes. However, when we select common continuous sequences of 60bp, the length of Agilent microarray probes, across human, rat and mouse orthologous genes, we find such sequences in only 2,280 (16.0%) of the 14,217 ortholog groups in the HomoloGene database [16]. For this reason, it has been thought difficult to make a microarray that can measure multiple-species using commercial arrays.

In order to design a microarray that can observe gene expression across multiple species, we here tried to relax the hybridization condition that requires a complete complementary sequence. We experimentally know that hybridization sometimes occurs even when a few mutations exist between two sequences. Therefore, microarray probes might be able to hybridize to target sequences even when they possess mutations. We therefore measured the effects of mutations by developing a microarray in which the probes have artificial mutations, and checked the number and position of mutations allowable for the measurement of expression levels on probe sequences.

We generated a microarray in which the probes had 104 types of artificial mutations (Section II-B). Fig. 3 shows the relations between mutation positions, numbers of mutations and expression levels observed. This figure indicates that the mutations affect the observed gene expression levels. In particular, expressions levels were found to decrease markedly with increases in the number of mutations in a probe. In addition, the mutation position had a large impact on the observed expression level. For example, even if five consecutive mutations existed in the 3'-end region, the probes could accurately measure expression, whereas the existence of a single mutation in the 5'-end regions of a probe resulted in a decrease in expression to below the expected level (Fig. 3).

These observations allowed us to relax the restrictions on primer design with regard to the need for identical consecutive 60-bp regions between all the target genes. To clarify the permissible number of mutations in the 3'-end region, we introduced a weighted distance measure to evaluate whether a probe could hybridize to each target or not.

B. Weighted distance between a probe and a gene

Existing research has relied upon used the existence of identical sequences between a probe and its target gene as a hybridization condition. Our artificial mutation experiments revealed that probes having specific mutations could still hybridize to the target genes. To measure whether the probe hybridizes to its target or not, we herein used a weighted distance that was reflected in the artificial mutation results (Fig. 1, Method).

The weighted distance between a probe and a 60-bp nucleotide sequence from a gene can be calculated as the summation of the weights on all the mutated positions. The weight is defined for each position in the probe. A mutation at a highly weighted position means that the mutation at that position has a large effect on hybridization, whereas a mutation at a lowly weighted position has only a small effect on hybridization. Distance between a probe and a gene is defined as the minimum weighted distance between the probe and nucleotide sequence from the gene. We could efficiently calculate the distances using a dynamic programming method. We determined that a probe could hybridize to its target gene when its distance is less than 10, while a probe cannot hybridize to genes when its distance is more than 20. In other words, a probe for an ortholog group is required to satisfy the set of conditions in which all distances between the probe and genes in the group are less than 10 and all distances between the probe and genes outside the group are more than 20. By changing the weighting on nucleotide positions and the hybridization thresholds, we can adjust the permissible hybridization conditions.

C. Design of the inter-species microarray

We here introduce a microarray probe design method that can measure gene expression across multiple species. In this paper, as an example, we present our design for a microarray that can measure human, rat and mouse gene expression to confirm the feasibility of our design strategy.

Generally, each probe on a microarray for a single species has to satisfy two conditions (Fig. 1(a)). One is that the probe can hybridize to the target genes, and the other is that the probe does not hybridize to any non-target genes. An extension of these conditions allows us to design probes for the array that can measure expression across multiple species, referred to as an inter-species array (Fig. 1(b)). Each probe on the inter-species array satisfies the above two conditions. For example, when an ortholog group contains three orthologous genes originating from humans, rats and mice, a probe for the group has to hybridize to all three genes. Further, the probe for an ortholog group must hybridize to no human, rat or mouse genes except those in the group. Probes designed according to this strategy have the potential to measure all orthologous genes. Furthermore, most of the probes can measure gene expression in species for the entire period from when the target species diverged on the evolutionary tree.

This set of conditions might appear too restrictive to design inter-species probes that can identify evolutional changes and entire cellular mechanisms. We calculated the number of possible probes and found inter-species probes for 6,683 out of 14,217 ortholog groups in the HomoloGene database [16]. Our target species are more diverse than those of Gilad *et al*. [13], but the number of designed probes is three times that used by Gilad *et al*. Furthermore, by limiting the inter-species array to only humans and rats, we can design probes for 6,792 ortholog groups (109 groups more than for the three-species array) and probes for 11,388 groups (4,705 groups more than the three-species array) could be generated for an inter-species array for rats and mice.

IV. Conclusions

We introduced a microarray that can be used to observe gene expression over multiple species with a single probe. We first showed that the presence of mutations close to the 3'-end of probes had very little effect on measurement of expression levels, while the presence of a single mutation close to the 5'-end of probes had a large effect on observed expression levels. We then introduced a microarray probe design method that finds probes hybridizing target orthologous genes but not hybridizing non-target genes. As the array operates simply by changing the probes on an Agilent commercial customized array, and conventional facilities and protocols can be used, the method is cost-efficient. We applied the method to measure human, rat and mouse expression levels with one microarray covering 6,683 genes (ortholog groups). The array could also be used to measure gene expression for distantly related species such as between humans and mice.

Acknowledgment

This work was partially supported by Grant-in-Aid for Scientific Research on Innovative Areas (23126504) from the Ministry of Education, Culture, Sports, Science and Technology, Japan. AO is currently with Institute of Genome Research, The University of Tokushima, Japan. JGC-S Scholarship to AO.

References

[1] Rhesus Macaque Genome Sequencing and Analysis Consortium, "Evolutionary and Biomedical Insights from the Rhesus Macaque Genome," *Science*, vol. 316, no. 5822, pp. 222–234, 2007.

[2] S. A. Rifkin, J. Kim, and K. P. White, "Evolution of gene expression in the Drosophila melanogaster subgroup," *Nat Genet*, vol. 33, no. 2, pp. 138–144, 2003.

[3] P. Khaitovich, W. Enard, M. Lachmann, and S. Pääbo, "Evolution of primate gene expression." *Nature Reviews Genetics*, vol. 7, no. 9, pp. 693–702, 2006.

[4] I. Yanai and C. P. Hunter, "Comparison of diverse developmental transcriptomes reveals that coexpression of gene neighbors is not evolutionarily conserved." *Genome Research*, vol. 19, no. 12, pp. 2214–2220, 2009.

[5] N. Irie and S. Kuratani, "Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis." *Nature Communications*, vol. 2, p. 248, 2011.

[6] A. T. Kalinka, K. M. Varga, D. T. Gerrard, S. Preibisch, D. L. Corcoran, J. Jarrells, U. Ohler, C. M. Bergman, and P. Tomancak, "Gene expression divergence recapitulates the developmental hourglass model," *Nature*, vol. 468, no. 7325, pp. 811–814, 2010.

[7] P. Khaitovich, I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Pääbo, "Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees," *Science*, vol. 309, no. 5742, pp. 1850–1854, 2005.

[8] J. M. Ranz, C. I. Castillo-Davis, C. D. Meiklejohn, and D. L. Hartl, "Sex-dependent gene expression and evolution of the Drosophila transcriptome." *Science*, vol. 300, no. 5626, pp. 1742–1745, 2003.

[9] S.-I. Morinaga, A. J. Nagano, S. Miyazaki, M. Kubo, T. Demura, H. Fukuda, S. Sakai, and M. Hasebe, "Ecogenomics of cleistogamous and chasmogamous flowering: genome-wide gene expression patterns from cross-species microarray analysis in Cardamine kokaiensis (Brassicaceae)," *Journal of Ecology*, vol. 96, no. 5, pp. 1086–1097, 2008.

[10] J. P. Hammond, H. C. Bowen, P. J. White, V. Mills, K. A. Pyke, A. J. M. Baker, S. N. Whiting, S. T. May, and M. R. Broadley, "A comparison of the Thlaspi caerulescens and Thlaspi arvense shoot transcriptomes," *New Phytol*, vol. 170, no. 2, pp. 239–260, 2006.

[11] Y. Lu, P. Huggins, and Z. Bar-Joseph, "Cross species analysis of microarray expression data." *Bioinformatics*, vol. 25, no. 12, pp. 1476–1483, 2009.

[12] A. Oshlack, A. E. Chabot, G. K. Smyth, and Y. Gilad, "Using DNA microarrays to study gene expression in closely related species," *Bioinformatics*, vol. 23, no. 10, pp. 1235–1242, 2007.

[13] Y. Gilad, S. A. Rifkin, P. Bertone, M. Gerstein, and K. P. White, "Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles," *Genome Research*, vol. 15, pp. 674–680, 2005.

[14] A. Mortazavi, B. A. Williams, K. Mccue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[15] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching *et al.*, "A gene atlas of the mouse and human protein-encoding transcriptomes," *Proc Natl Acad Sci*, vol. 101, no. 6, pp. 6062–6067, 2004.

[16] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese *et al.*, "Database resources of the National Center for Biotechnology Information." *Nucleic Acids Research*, vol. 34, no. Database issue, pp. D173–80, 2006.

[17] D. Gusfield, *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, 1st ed. Cambridge University Press, 1997.

[18] S. Rozen and H. J. Skaletsky, "Primer3 on the WWW for general users and for biologist programmers," in *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, K. S and M. S, Eds. steverozen.net, 2000, pp. 365–386.