Imprinted gene detection in Arabidopsis thaliana

Reiner Schulz **(reiner.schulz@kcl.ac.uk)**
Dept. of Medical and Molecular Genetics, King's College London
Sushma Tiwari, Rod Scott **({s.m.tiwari,bssrjs}@bath.ac.uk)**
Dept. of Biology and Biochemistry, University of Bath
Rebecca Oakey **(rebecca.oakey@kcl.ac.uk)**
Dept. of Medical and Molecular Genetics, King's College London

Abstract - *We use the self-organizing map (SOM) to detect novel imprinted genes in A.thaliana from microarray gene expression measurements in several interploidy and mutant A.thaliana crosses. In sexually reproducing organisms, the transcription of the maternally as well as the paternally inherited alleles of a gene is the rule. Imprinted genes are an exception to this rule. Their number is relatively small, but they are often involved in the regulation of growth, particularly seed growth in plants. Only five genes are known to be imprinted in A.thaliana, but given that many more are known in other species, A.thaliana likely harbors additional imprinted genes. The self-organizing map is a suitable tool for the identification of novel imprinted genes. The map projections of known imprinted genes yield sets of genes with similar expression profiles that are probable candidates for imprinting. Assuming a gene is imprinted, the its expression measurements across the different crosses are expected to follow a certain order. The number of deviations from this expected order provide a means to rank the SOM-generated candidates and thus, to prioritize their verification in the laboratory.*

Keywords - bioinformatics; self-organizing map; microarray; Arabidopsis; imprinting

1 Introduction

Genetic imprinting is a form of epigenetic gene regulation that makes use of heritable molecular markers (imprints) which are attached to, but not part of the genome. They convey information about the parental origin of alleles to the offspring, and, depending on the gene, restrict its expression (transcription into mRNA) to only either the maternally or the paternally inherited alleles. Imprinted genes of the former category are called *maternally expressed*, while the latter complementary category comprises the *paternally expressed* imprinted genes. Nuclear transfer experiments in mice lead to the discovery of the imprinting phenomenon [12, 19]. Since then, more than 70 and 40 imprinted genes have been identified in mouse [5] and human [14], respectively. The total number of imprinted genes among the roughly 25,000 genes in either genome is estimated to range between 100 and 300. In *A.thaliana*, the model genetic organism for the plant kingdom [15], only five genes, *MEA*, *FWA*, *FIE*, *FIS2* and *PHE* are known to be imprinted or show signs of imprinting [4, 8, 7, 10]. In mammals and plants, imprinted genes are disproportionately often involved in the regulation of growth. In *A.thaliana*, all known imprinted genes are associated with seed growth and development so that the discovery of novel imprinted genes is potentially valuable.

The two parts of the seed that result from double fertilization are the embryo and the endosperm. The diploid embryo is the product of the haploid egg fusing with one of the two haploid sperm cells from the pollen grain. The other sperm fuses with the two polar nuclei of the central cell to form the triploid endosperm, whose purpose is to feed the embryo during germination. So, under normal circumstances, each embryonic cell carries one maternal (from the egg) and one paternal (from the sperm) allele per gene, while a cell of the endosperm hosts two maternal (from the polar nuclei) alleles and one paternal (from the sperm) allele.
Fertile tetraploid and hexaploid *A.thaliana* plants can be generated and crossed with normal diploid plants (interploidy crosses) to produce seeds with either excess maternal or excess paternal genomes [16]. In addition, 'paternalization' of the polar nuclei ($\Rightarrow$ excess of paternal genomes) and 'maternalization' of the sperm ($\Rightarrow$ excess of maternal genomes) can be achieved in 'mutant' crosses involving *A.thaliana* plants with either an antisense RNA down-regulated *MET* gene (*METas*) or a mutation in the *MEA* gene (*mea*) [1, 18]. Given an imprinted gene, its transcript's relative abundance when compared to normal (wildtype) seeds ought to change across the different interploidy and mutant crosses in a manner that distinguishes it from non-imprinted genes.
Expression microarrays provide the means to quantitatively measure the amount of transcript per gene for many genes at a time [11, 17]. Here, we used microarrays to create an expression profile for each of thousands of *A.thaliana* genes across four interploidy and three mutant crosses. A gene's expression profile corresponds to a location in a high-dimensional feature space where each dimension represents a different microarray experiment.
The self-organizing map (SOM) lends itself to the analysis and visualization of large sets of high-dimensional feature vectors [9]. It performs a non-linear projection of the feature space onto a typically two-dimensional regular lattice of nodes, the map or map space. Each node is associated with a location in feature space. The set of all nodes give rise to a tessellation of the feature space into Voronoi cells. The data-driven, unsupervised training of the SOM incrementally moves the nodes in feature space so that their post-training positions tend to approximate a manifold that is representative of the training data, and so that iff two nodes are adjacent on the map, their Voronoi cells tend to be adjacent. Intuitively, this means that after training, the closer two nodes are on the map the more similar regions of the feature space they represent.
The SOM has been used previously in microarray data analysis [20], but has not been applied toward the discovery of imprinted genes, whose distinct expression profiles are likely to separate them on the map from the bulk of non-imprinted genes. Specifically, the nodes of the map that come to represent known imprinted genes will likely also represent other genes with similar expression profiles, which suggest that they too may be imprinted.
Due to the parent-of-origin-specific transcription of imprinted genes and the varying ratios of maternal versus paternal genomes in the selected *A.thaliana* crosses, the relative expression levels of imprinted genes are expected to follow a particular order across the different crosses. This allows a ranking to be superimposed on a SOM-generated set of imprinting candidates where a candidate ranks the higher the better it adheres to the expected order. The verification of imprinting by molecular biology techniques in the laboratory is time-consuming and expensive. The ranking, in addition to providing a test of the SOM's results, helps to focus the biological verification on the most promising candidates.
The remainder of this paper details the SOM-based methodology for the detection of imprinted genes in *A.thaliana* and reports first results. The emphasis is on the data analysis

with the SOM and the mathematical model that underlies the ranking of SOM-generated candidates, rather than biological sample preparation and data generation.

2 Materials and Methods

Two sets of microarray experiments were conducted, each using different plant material and array technology. For the first set, wildtype seeds (2x X 2x cross) and seeds of four interploidy (6x X 2x, 4x X 2x, 2x X 4x, 2x X 6x) and three mutant (*mea* X 2x, *METas* X 2x, 2x X *METas*) crosses were harvested five days after pollination (5dap). Separately for each cross, mRNA samples were extracted from whole seeds and prepared for hybridization to Agilent 22K two-dye (Cy3 and Cy5) arrays [3], each carrying Ceres-designed (`http://www.ceres-inc.com`) oligonucleotide probe sets. For each non-wildtype cross, one array was hybridized using a Cy3-labeled sample together with a Cy5-labeled wildtype sample. The dyes were then swapped and the experiment repeated on a second array, for a total of 14 experiments. Each array produced two expression measurements per probe set: s_{ij} for the non-wildtype and s'_{ij} for the wildtype sample where i and j identify the array and the probe set, respectively. All subsequent analyses were based on the log-ratio ($\log_2(s_{ij}/s'_{ij})$) of these values. Probe sets whose measurements were deemed unreliable in one or more experiments were excluded from further analysis, leaving 19,109 probe sets with log-ratios for all 14 experiments.

The second set of microarray experiments used seeds from the same wildtype and interploidy crosses, but harvested at 3dap. Total mRNA samples made from the seeds were sent to NASC (`http://affy.arabidopsis.info`) where they were processed and hybridized to Affymetrix ATH1 arrays (five arrays, one per cross) [2]. The results of each non-wildtype experiment were compared to the wildtype experiment using the pairwise expression analysis feature of Affymetrix' GCOS v1.2. GCOS computes a signal log ratio (SLR) for each probe set, whose definition approximately equals that of the log-ratio in the first set of experiments. The second set of experiments thus produced an additional four log-ratios per probe set.

The correspondence between the Ceres and Affymetrix probe sets was established via the TAIR/AGI and GenBank accession numbers of the represented genes in the Ceres and Affymetrix annotation. 11,636 genes were represented by both Ceres and Affymetrix probe sets and had their expression measured successfully in all 18 experiments.

Two filters were applied to eliminate probe sets whose log-ratios did not change significantly across experiments. Treating the two dye-swapping experiments per non-wildtype cross as replicates, 6,859 Ceres probe sets were differentially expressed in at least one experiment (one-way ANOVA; $\alpha = .05$). Similarly, 9,399 Affymetrix probe sets showed a difference in expression relative to the wildtype experiment in one or more non-wildtype experiments (GCOS change call $\neq$ 'no change').

2,547 genes (among them *PHE* and *MEA*) were represented by probe sets on both types of array, were measured successfully in all 18 experiments, and passed both filters. The 18-dimensional vectors composed of the these genes' log-ratios constituted the training data for the SOM. Data preprocessing, SOM training and analysis were performed using the SOM toolbox v2.0β [21]. The 18 data dimensions were independently normalized to each have a mean of zero and a variance of one. The training parameters were set based on high-level arguments given to the training function `som_make`. The ratio of the data's two largest eigenvalues determined the map's size of 18 by 14 nodes. The SOM's weights were either initialized to form a regular grid in the plain spanned by the data's two largest eigenvectors and then trained for four epochs using the batch algorithm (deterministic training), or the

Table 1: Model of the microarray-measured signal & expected order among the signal ratios

Cross	Embryo	Endo-sperm	Rest	Normalized Signal	Signal Ratios, Expected Order and Assumptions
2 x 2	$1m+1p$	$2m+1p$	2	$\frac{\mathrm{Em}\,(m+p)+\mathrm{En}\,(2\,m+p)+2\,\mathrm{Re}}{2\,\mathrm{Em}+3\,\mathrm{En}+2\,\mathrm{Re}}$	1
				preferentially expressed allele →	maternal | paternal
6 x 2	$3m+1p$	$6m+1p$	6	$\frac{\mathrm{Em}\,(3\,m+p)+\mathrm{En}\,(6\,m+p)+6\,\mathrm{Re}}{4\,\mathrm{Em}+7\,\mathrm{En}+6\,\mathrm{Re}}$	6x2/2x2
					$>_{m>p}$ | $<_{m=0,p=1}$
4 x 2	$2m+1p$	$4m+1p$	4	$\frac{\mathrm{Em}\,(2\,m+p)+\mathrm{En}\,(4\,m+p)+4\,\mathrm{Re}}{3\,\mathrm{Em}+5\,\mathrm{En}+4\,\mathrm{Re}}$	4x2/2x2
					$>_{m>p}$ | $<_{m=0,p=1}$
2 x 4	$1m+2p$	$2m+2p$	2	$\frac{\mathrm{Em}\,(m+2\,p)+\mathrm{En}\,(2\,m+2\,p)+2\,\mathrm{Re}}{3\,\mathrm{Em}+4\,\mathrm{En}+2\,\mathrm{Re}}$	2x4/2x2
					$>_{m>p}$ | $<_{m=0,p=1}$
2 x 6	$1m+3p$	$2m+3p$	2	$\frac{\mathrm{Em}\,(m+3\,p)+\mathrm{En}\,(2\,m+3\,p)+2\,\mathrm{Re}}{4\,\mathrm{Em}+5\,\mathrm{En}+2\,\mathrm{Re}}$	2x6/2x2
					$>_{\substack{m=1,p=0,\\ \mathrm{Em}+\mathrm{En}>\mathrm{Re}}}$ | $<_{m=0,p=1}$
mea x 2	$0m+2p$	$0m+3p$	2	$\frac{2\,\mathrm{Em}\,p+3\,\mathrm{En}\,p+2\,\mathrm{Re}}{2\,\mathrm{Em}+3\,\mathrm{En}+2\,\mathrm{Re}}$	*mea*x2/2x2
2 x *METas*	$2m+0p$	$3m+0p$	2	$\frac{2\,\mathrm{Em}\,m+3\,\mathrm{En}\,m+2\,\mathrm{Re}}{2\,\mathrm{Em}+3\,\mathrm{En}+2\,\mathrm{Re}}$	2x*METas*/2x2
					$>_{m>p}$ | $<_{m<p}$
METas x 2	$0m+2p$	$0m+3p$	2	$\frac{2\,\mathrm{Em}\,p+3\,\mathrm{En}\,p+2\,\mathrm{Re}}{2\,\mathrm{Em}+3\,\mathrm{En}+2\,\mathrm{Re}}$	*METas*x2/2x2

weights were randomly initialized and trained for 10,248 epochs using the sequential algorithm (probabilistic training). In the latter case, training was repeated 100 times, each time with a different initial random number generator seed.

After training, the map projection of *PHE*, that is, the node closest to *PHE* in feature space, defined the set of *strictly PHE-similar* genes, i.e., all genes that were closer to *PHE*'s map projection than to any other node. Analogously, *MEA*'s map projection defined the set of *strictly MEA-similar* genes. With probabilistic training, these sets varied from SOM to SOM due to different initial random conditions. For comparison, supersets of *loosely similar* genes were created based on a post-training hierarchical clustering of the nodes (cluster seeds: local minima of the inter-node distance matrix; Ward inter-cluster distance metric). The map projections of *PHE* and *MEA* each became part of a cluster composed of multiple nodes whose union defined the set of genes loosely similar to *PHE* and *MEA*, respectively.

The genes in the strictly and loosely similar sets were ranked according to how well their log-ratios adhered to the order that is expected under the assumption of preferential paternal expression in the case of *PHE*-similar genes and preferential maternal expression for *MEA*-similar genes. The expected orders derive from a mathematical model of the microarray-measured signal and are consistent with observations on seed size in the different *A.thaliana* crosses and the conflict theory [13].

The model, summarized in Tab. 1, expresses the signal measured for a probe set in terms of the contributions of the maternally and paternally derived alleles and the different tissues of the seed. For example, consider the wildtype cross (top row of Tab. 1). The seed is composed of embryo, endosperm and a rest (seed wall, etc.). For each gene, an embryonic cell carries two alleles, one maternal and one paternal ($1m+1p$) and an endosperm cell hosts three alleles,

two maternal and one paternal ($2m + 1p$). A cell from the remaining, entirely maternally derived tissues contains two alleles that are considered sexless (2 vs. $2m$) since these tissues are mature, while the effects of imprinting are mostly confined to early development.
The parameters $m \geq 0$ and $p \geq 0$ reflect the relative mRNA contributions of the maternal and paternal alleles, respectively. For a non-imprinted gene $m = p = 1$, that is, all alleles are active. For an imprinted gene, if it is exclusively paternally expressed then $m = 0$ and $p = 1$, and correspondingly, $m = 1$ and $p = 0$ in case of exclusive maternal expression.
Embryo, endosperm and rest each contribute a different amount of mRNA to the sample, modeled by the coefficients $\mathrm{Em} \geq 0$, $\mathrm{En} \geq 0$ and $\mathrm{Re} \geq 0$ that are assumed constant across the different crosses. Thus, the total mRNA for a gene in the wildtype sample is modeled as $\mathrm{Em}\ (m + p) + \mathrm{En}\ (2\,m + p) + 2\,\mathrm{Re}$ (numerator in column 'Normalized Signal').
The measured signal for a probe set is normalized by the global average fluorescence of the array. The model assumes that the average fluorescence is proportional to the total genome-wide amount of mRNA in the sample, and that this amount is proportional to the sum over all seed tissues (embryo, endosperm, rest) of the weighted number of genomes per cell, i.e., for the wildtype sample, $2\,\mathrm{Em} + 3\,\mathrm{En} + 2\,\mathrm{Re}$ (divisor in column 'Normalized Signal'). Hence, the model does not take into account saturation, compensation and other non-linear effects.
The last column in Tab. 1 indicates how, under certain assumptions, the ratios of the normalized signals for the non-wildtype crosses versus the wildtype cross are expected to be ordered. For example, given an exclusively paternally expressed imprinted gene ($m = 0$, $p = 1$), 6x2/2x2 is expected to be less than 4x2/2x2. The opposite is true for a maternally expressed imprinted gene, even if one only assumes preferential maternal expression ($m > p$).
A gene's adherence to the expected order was quantified in terms of the number of pairwise comparisons between the gene's signal ratios that violated the expected order [6]. This was done separately for the two sets of experiments since the signal ratios were not directly comparable across array technologies. Below, E refers to the sum of these separate counts.

3 Results

Fig. 1 shows that the deterministically trained SOM projected *PHE* to a node near one of the maps' corners. In total, 15 genes were mapped to this node. They compose this SOM's set of strictly *PHE*-similar genes. The node constituted a local minimum in the intra-node distance matrix and hence, served as a seed in the subsequent hierarchical clustering of the SOM's nodes. The cluster that grew from this seed contained 101 genes, the set of loosely *PHE*-similar genes (see Fig. 1B). The extreme and isolated map position of *PHE* suggests that *PHE* belongs to a relatively small and distinguished set of genes with similarly exceptional expression patterns. *PHE* being at a local minimum in the distance matrix further indicates that *PHE* is representative of this set, which increases the likelihood of finding truly imprinted genes among the *PHE*-similar genes.
MEA and seven other strictly similar genes were projected to a node closer to the map's center. Unlike for *PHE*, the node did not belong to a neighborhood that was isolated from the other nodes. This indicates that *MEA*'s expression profile does not distinguish it significantly from the bulk of genes. The node for *MEA* did not constitute a cluster seed and was among the last to be assigned to a cluster. Its peripheral position within the cluster (see Fig. 1B) points to *MEA* not being representative of the corresponding set of 263 loosely *MEA*-similar genes.
The results obtained with the probabilistically trained SOMs confirmed that *PHE* is representative of a distinct set of genes, while *MEA* lacks this property. Like in Fig. 1, *PHE* was

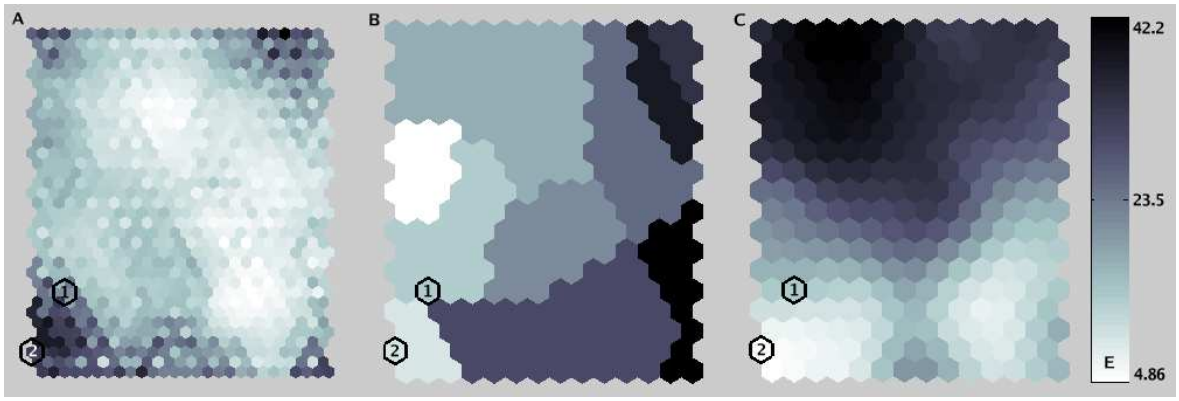

Figure 1: Map visualizations of the deterministically trained SOM. **A** Inter-node distance matrix. The brighter an element the closer the corresponding adjacent nodes are to one another in feature space. *PHE* (2) is projected near the bottom-left corner of the map. Dark elements isolate it from the rest of the map, suggesting that *PHE* belongs to a set of unusually expressed genes, placed far from the bulk of genes in feature space. *MEA* (1) is closer to the map's center and less isolated, i.e., *MEA* is not obviously part of a distinguished set of genes. **B** Hierarchical clustering result. Each final cluster is shaded differently. *PHE* and *MEA* belong to different but adjacent clusters. *PHE* is representative of its cluster for which it was the seed, while *MEA*'s position at the border with two other clusters indicates that it is not representative of a particular set of genes. **C** Distribution of E (assuming paternal expression) across the map. *PHE* is located in the region with the lowest E values, while *MEA* is in a region of roughly average E values, consistent with $E_{PHE} = 0$ and $E_{MEA} = 27$.

always projected to a corner of the map and *MEA* to a more central location. Fig. 2A shows that in general, the probabilistic training results for *PHE* were much more robust in the face of changing initial random conditions than those for *MEA*. For example, 15 genes were each assigned to the strictly *PHE*-similar set by at least 2/3 of the probabilistic SOMs. These 15 genes corresponded exactly to the strictly *PHE*-similar set of the deterministic SOM. In general, probabilistic and deterministic training results for *PHE* corresponded much more closely than for *MEA* (data not shown).

On average, a gene in the SOM training set (2,547 genes) violated the order among the signal ratios that is expected of an imprinted gene 24.6 times in the case of paternal expression and 25.4 times in the case of maternal expression. This is consistent with the perchance expected average of 25 in both cases (the theoretical maximum is 50). *PHE* (paternally expressed) adhered to the expected order ($E_{PHE} = 0$). Hence, the *PHE*-similar genes were expected to respect the order too. *MEA* (maternally expressed) did not adhere to the expected order ($E_{MEA} = 27$), and so, it was unlikely that *MEA*-similar genes would respect the order. Figures 1C and 2B illustrate that these expectations were mostly met. With probabilistic training, $\bar{E}$ over the sets of *PHE*-similar genes was always less than half the overall average of 24.6. For the 15 genes that were judged strictly *PHE*-similar by both the deterministic SOM and 2/3 of the probabilistic SOMs, $\bar{E}$ was particularly low at 4.4.

For comparison, a k-means clustering of the SOM training data was performed ($k = 86$; minimized the Davies-Bouldin index among $k = 1, \ldots, 102$). The cluster that contained *PHE* composed 16 genes in total, 14 of which were among the above 15 strictly *PHE*-similar genes.

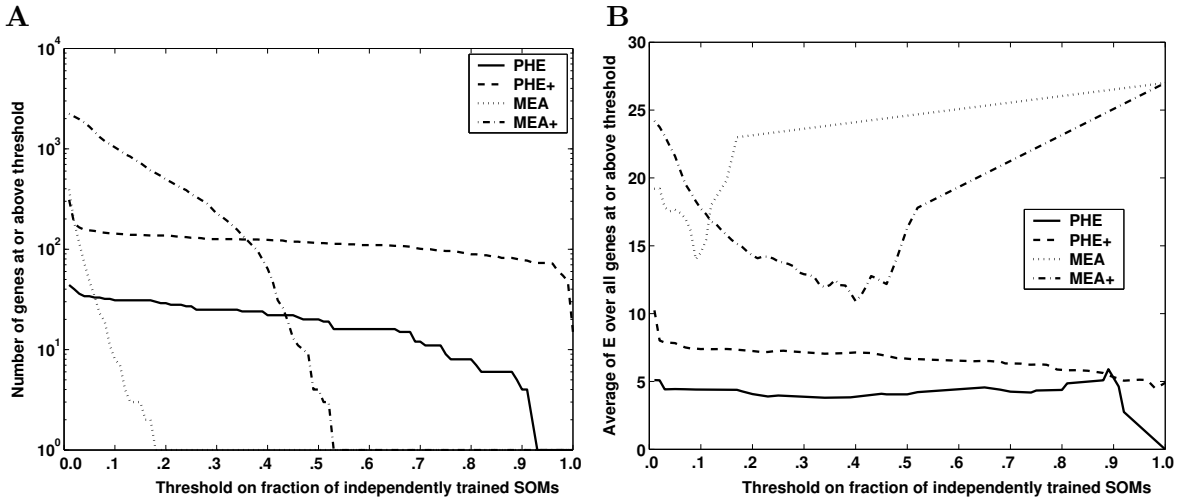

Figure 2: **A** The number of genes (logarithmic y-axis) that were put in the strictly and loosely similar sets for *PHE* (solid and dashed curves) and *MEA* (dotted and dash-dotted curves), respectively, by at least a given fraction of the 100 independently trained SOMs (x-axis). The numbers decreased at a much higher rate for *MEA* than for *PHE*, meaning that the composition of the *PHE*-similar sets was less dependent on the initial random conditions of training than in the case of *MEA*. **B** The average of E ($\bar{E}$; y-axis) for the genes strictly and loosely similar to *PHE* and *MEA*, respectively, in at least the given fraction of independently trained SOMs (x-axis). For *PHE* with $E_{PHE} = 0$, $\bar{E}$ is expected to be lower the larger the required fraction of SOMs (x). This largely matched our observations, except for $.4 \leq x \leq .9$ where $\bar{E}$ tended to increase slightly for strictly *PHE*-similar genes. As expected, due to $E_{MEA} = 27$, $\bar{E}$ for *MEA*-similar genes was always greater than for *PHE*.

4 Conclusions

The SOM detected 14 particularly promising novel imprinting candidate genes, based on being expressed similarly to the known imprinted gene *PHE* in seven distinct *A.thaliana* crosses. Variations in the training algorithm or the initial random conditions of training had little impact on this result. The candidates adhered well to the predictions of an independent biologically-grounded mathematical model of microarray-measured expression, and all but one gained further support from the results of a k-means clustering analysis. The candidates' true imprinting status is currently being determined in the laboratory, with priority given to those candidates that best agree with the model's predictions.

For *MEA*, the second known imprinted gene in the training set, the composition of the sets of similarly expressed genes much depended on the conditions of SOM training. *MEA*'s expression profile was not distinguished enough to yield robust results, perhaps partially due to one of the crosses involving a mutated *MEA* gene.

The SOM's use was motivated also by its visualization abilities, distinguishing it from other clustering techniques. The maps in Fig. 1 and maps of all feature space dimensions (microarray experiments) were turned into a web-accessible graphical interface to the data, supporting its visual exploration and queries of the underlying database to retrieve detailed annotation. For example, we were able to see that the multiple experiments for the same cross differing only in array technology, developmental time point, or labeling dye produced similar maps.

ACKNOWLEDGMENT: We thank Ceres, Inc., for providing the first set of microarray data.

References

[1] S. Adams, et al. Parent-of-origin effects on seed development in arabidopsis thaliana require DNA methylation. *Development*, 127:2493–502, 2000.

[2] Affymetrix, Inc. *Expression Analysis Technical Manual*. 2004. `http://www.affymetrix.com/Auth/support/downloads/manuals/expression_ever_manual.zip`.

[3] Agilent Technologies, Inc. *Agilent 60-mer oligo microarray processing protocol*. 2004. `http://www.chem.agilent.com/scripts/literaturePDF.asp?iWHID=33686`.

[4] C. Baroux, et al. Genomic imprinting during seed development. *Adv Genet*, 46:165–214, 2002.

[5] C. Beechey, et al. *Mouse Imprinting Data and References*. MRC Mammalian Genetics Unit, Harwell, 2005. `http://www.mgu.har.mrc.ac.uk/research/imprinting/`.

[6] A. Jonckheere. A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41:133–45, 1954.

[7] T. Kinoshita, et al. One-way control of FWA imprinting in arabidopsis endosperm by DNA methylation. *Science*, 303(5657):521–3, 2004.

[8] C. Kohler, et al. The polycomb-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene PHERES1. *Genes Dev*, 17:1540–53, 2003.

[9] T. Kohonen. *Self-Organizing Maps*. Springer, 3rd edition, 2001.

[10] M. Luo, et al. Expression and parent-of-origin effects for FIS2, MEA, and FIE in the endosperm and embryo of developing Arabidopsis seeds. *Proc Natl Acad Sci USA*, 97(19):10637–42, 2000.

[11] U. Maskos and E. Southern. Oligonucleotide hybridizations on glass supports. *Nucleic Acids Res*, 20(7):1679–84, 1992.

[12] J. McGrath and D. Solter. Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, 37:179–83, 1984.

[13] T. Moore and D. Haig. Genomic imprinting in mammalian development: a parental tug-of-war. *Trends Genet*, 7(2):45–9, 1991.

[14] I. Morison, et al. The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res*, 29(1):275–6, 2001. `http://www.otago.ac.nz/IGC`.

[15] S. Rhee, et al. The Arabidopsis Information Resource (TAIR). *Nucleic Acids Res*, 31(1):224–8, 2003.

[16] R. Scott, et al. Parent-of-origin effects on seed development in Arabidopsis thaliana. *Development*, 125(17):3329–41, 1998.

[17] E. Southern. DNA chips: analysing sequence by hybridization to oligonucleotides on a large scale. *Trends Genet*, 12(3):110–5, 1996.

[18] M. Spielman, et al. The epigenetic basis of gender in flowering plants and mammals. *Trends Genet*, 17:705–711, 2001.

[19] A. Surani, et al. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature*, 308:548–50, 1984.

[20] P. Tamayo, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA*, 96(6):2907–12, 1999.

[21] J. Vesanto, et al. Self-organizing map in Matlab: the SOM toolbox. In *Proc Matlab DSP Conference 1999*, pp. 35–40, 1999. `http://www.cis.hut.fi/projects/somtoolbox/`.