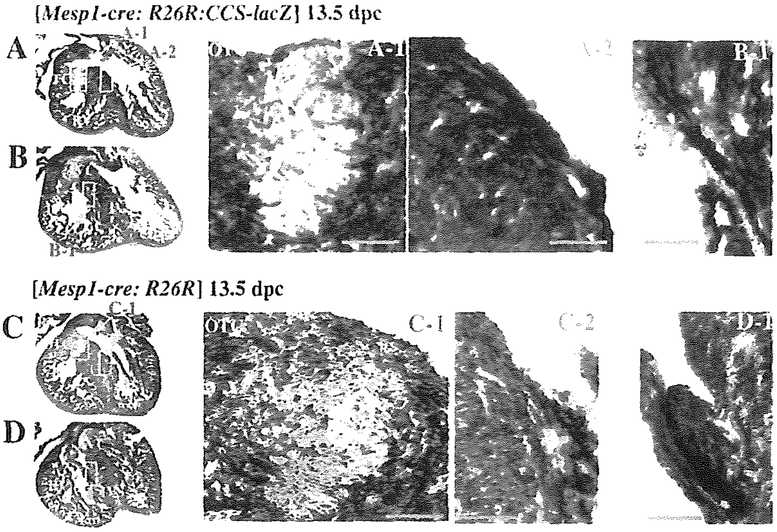
[Mesp1-cre: R26R:CCS-lacZ] 13.5 dpc
A
A-1
A-2
B
B-1
[Mesp1-cre: R26R] 13.5 dpc
C
C-1
C-2
D
D-1
OT

Fig. 3.

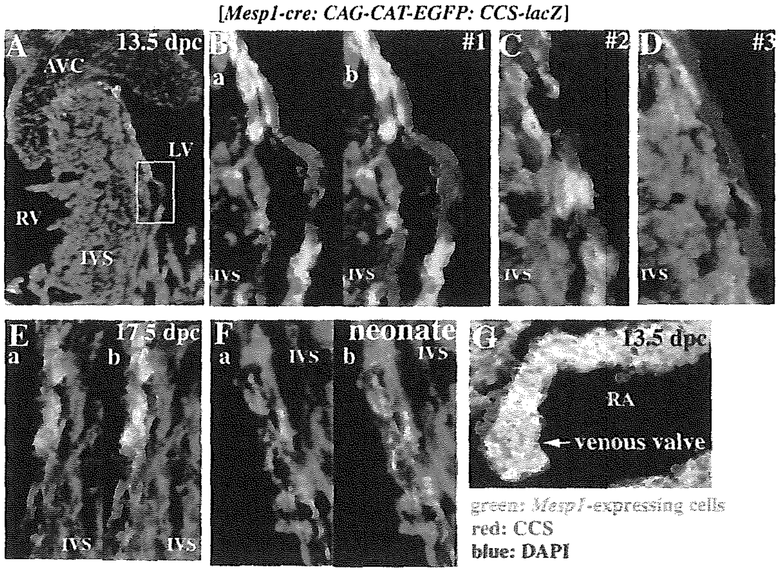

Fig. 4.

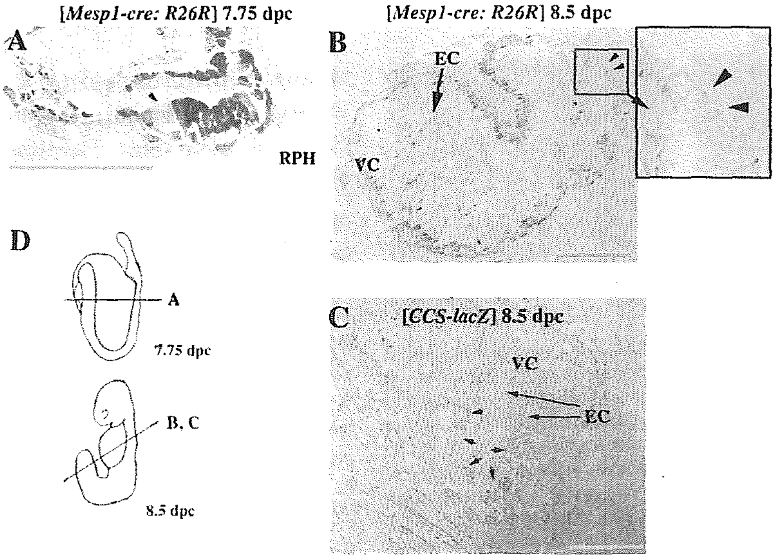

Fig. 5.

Fig. 3. Comparison of β-gal staining patterns between *Mesp1-cre:R26R:CCS-lacZ* triple hetero-embryos and *Mesp1-cre:R26R* embryos. **A,B:** Sections of the heart of the *Mesp1-cre: R26R:CCS-lacZ* embryo at 13.5 days post coitum (dpc). **C,D:** Compared with the *Mesp1-cre: R26R* embryo (boxes in C-2 and D-1), *Mesp1*-nonexpressing cells along the interventricular (IVS) were barely observed in the *Mesp1-cre: R26R:CCS-lacZ* triple hetero-embryos (A,B; boxes in A-2 and B-1). However, *Mesp1*-nonexpressing cells in outflow tract (OT) cushion regions were also detected, even in the triple hetero-embryos (box A-1). Original magnification, ×100. The images in the boxed area are magnified. Sectioning planes of A, B, and C, D are the same as those illustrated in Figure 1E: B and C, respectively. Sections were counterstained with eosin. AVC, atrioventricular cushion; IVS, interventricular septum; RA, right atrium; RV, right ventricle; Scale bar = 100 μm.

Fig. 4. *Mesp1*-nonexpressing cells contribute to a subset of the ventricular cardiac conduction system (CCS). Triple immunostaining for *Mesp1*-expressing cells (green fluorescent protein [GFP] -positive cells; green), cells of the ventricular CCS (LacZ-positive cells; red), and nuclei (4′,6′-diamidino-2-phenylindole [DAPI] staining; blue) in a *Mesp1-cre:CAG-CAT-EGFP:CCS-lacZ* embryo. All images shown are merged views, and double immunostaining of GFP and LacZ (a) and a triple immunostaining image with additional DAPI staining (b) are shown in some cases. **A–D:** A merged view of the interventricular (IVS) region at 13.5 days post coitum (dpc). The boxed area of the ventricular CCS in A was magnified as shown in B. Other sections derived from additional embryos are shown in C and D. The presence of red cells suggests that *Mesp1*-nonexpressing cells actually belong to the ventricular CCS, whereas some *Mesp1*-expressing cells also colocalize here (yellow). Typical images of mixed cell populations are shown in B and C, whereas a red cell-dominant section is shown in D. Original magnification, ×400, except for A, which is ×100. **E,F:** Merged view in the IVS region in an embryo at 17.5 dpc (E) or in a neonate (F). Original magnification, ×400. Red cells (i.e., the *Mesp1*-nonexpressing cells belonging to the CCS) were observed even at later stages beyond 13.5 dpc. **G:** The region of the venous valves, which are proposed remnants of the embryonic sinoatrial (SA) ring, at 13.5 dpc. Almost all of the cells in this region were stained yellow, suggesting that the cells belonging to the venous valves are *Mesp1*-expressing. Original magnification, ×200. Sectioning planes are those between B and C, illustrated in Figure 1E. AVC, atrioventricular cushion; LV, left ventricle; OTC, outflow tract cushion; RA, right atrium; RV, right ventricle.

Fig. 5. Comparison of transverse sections of β-gal stained *Mesp1-cre:R26R* embryos with *CCS-lacZ* embryos at an earlier stage. **A:** At 7.75 days post coitum (dpc) in *Mesp1-cre:R26R* embryos, we observed a few *Mesp1*-nonexpressing cells within the primitive heart tube (arrow head). Original magnification, ×400. **B:** At 8.5 dpc, the regions of the *Mesp1*-nonexpressing cells in *Mesp1-cre:R26R* embryos, were observed more clearly (arrowheads). **C:** The β-gal-positive regions (i.e., the cells belonging to the CCS) were observed mainly in the subendocardial myocardium of 8.5 dpc *CCS-lacZ* mouse (arrows). Original magnification, ×200. **D:** Sectioning planes are illustrated. Sections were counterstained with eosin. EC, endocardium; RPH, right primitive heart tube; VC, ventricular chamber. Scale bars = 100 μm.

2000). There were only minimal contributions by neural crest cells in the AV cushions, as predicted by the β-gal activity in the *Mesp1-cre:R26R* mouse (Fig. 2C). Importantly, however, neural crest-derived mesenchyme was not observed in either part of the ventricle or the IVS (Fig. 2B), where *Mesp1*-nonexpressing cells were visible (Fig. 1B). This finding indicates that other cell types must contribute to this particular region. Intriguingly, the distribution of *Mesp1*-nonexpressing cells resembled that of the AVB and bundle branches and also the Purkinje fibers of the CCS. This prompted us to speculate that ventricular CCS cells might be derived from lineages that are distinct from both the neural crest and *Mesp1*-expressing mesodermal cells.

Mesp1-Nonexpressing Cells Contribute to the CCS

As a preliminary approach to determine whether or not *Mesp1*-nonexpressing cells did in fact reside in the CCS, we compared these cells with the β-gal expression patterns in embryonic hearts of *CCS-lacZ* transgenic mice. In these mice, the specialized CCS can be visualized by β-gal activity (Rentschler et al., 2001). In 13.5 dpc hearts from these transgenic animals, strong β-gal activity could be observed in part of the atrium, which could correspond to the SA node. This high level of activity could also be detected along the IVS, which demarcates the ventricular CCS, including the AVB and bundle branches (Fig. 2D–F). When comparing these results with those shown in Figure 1, the portion of the *Mesp1*-nonexpressing cell population along the IVS was found to show a similar pattern to the β-gal–positive regions in the *CCS-lacZ* mice, suggesting that these *Mesp1*-nonexpressing cells contribute to the ventricular CCS.

To provide direct evidence for our hypothesis that cells of the ventricular CCS are indeed derived from *Mesp1*-nonexpressing cells, we generated triple transgenic *Mesp1-cre:R26R:CCS-lacZ* mice. Because both the *CCS-lacZ* and *R26R* transgenic mice use β-gal as a marker, the entire region contributed by the *Mesp1*-nonexpressing cells in the IVS would become β-gal–positive in the triple hetero-embryonic hearts if our contention was correct. As shown in Figure 3, this was found to be the case, as all of the cells in the IVS had β-gal activity, which was in contrast to the corresponding sections of the *Mesp1-cre:R26R* embryo (Fig. 3C,D). Moreover, the region of the OT cushions had little β-gal activity even in the triple hetero-embryo (Fig. 3A), supporting our conclusion that this region is occupied mainly by cells of neural crest origin. Hence, these data suggest that the *Mesp1*-nonexpressing cells in the IVS belong to the ventricular CCS.

It was still unclear, however, whether all of the ventricular CCS is derived from *Mesp1*-nonexpressing cells, because both the *CCS-lacZ* and *R26R* reporter mice use the same β-gal marker. We, therefore, performed a similar series of studies using the *CAG-CAT-EGFP* strain (Kawamoto et al., 2000), in which GFP expression is dependent upon cre-mediated recombination and representative results are shown in Figure 4. *Mesp1*-nonexpressing cells at 13.5 dpc do indeed reside within the ventricular CCS (*Mesp1*-nonexpressing/*CCS-lacZ*-positive red cells in Fig. 4A–D), although it is clear that the CCS is also observed in the *Mesp1*-expressing cell populations (i.e., *Mesp1*-expressing/*CCS-lacZ*–positive yellow cells). In addition, after 4′,6′-diamidino-2-phenylindole (DAPI) staining, we observed that all of the green fluorescent protein (GFP) -negative *Mesp1*-nonexpressing cells belonged to the lacZ-positive cells of the ventricular CCS, because cells positive for DAPI alone (blue) were rarely observed along the IVS (b in Fig. 4A–C). To demonstrate the heterocellular origin of CCS more unequivocally and to analyze the ratios quantitatively, we generated serial sections of the embryonic heart along the anteroposterior axis and analyzed the staining patterns.

A total of three embryos were sectioned and 58 sections containing CCS-LacZ staining in the IVS region were further subjected to semiquantitative analysis (Supplementary Figure S1, which can be viewed at http://www.interscience.wiley.com/jpages/1058-8388/suppmat). However, as the CCS distributes peripherally in the IVS with multiple branchings, it is very difficult to quantify. We, therefore, roughly estimated the ratio by counting DAPI stained nuclei in each cell type and selected 28 typical sections, from which 16 showed a colocalization pattern for yellow and red cells (Fig. 4B,C). Of these 16 sections, 2 and 5 showed a red cell- and a yellow cell-dominant pattern, respectively (Fig. 4D, and data not shown). We have estimated that approximately 20% of the ventricular CCS, along the IVS, corresponds to *Mesp1*-nonexpressing cells. Moreover, red cells (i.e., the *Mesp1*-nonexpressing cells belonging to the ventricular CCS) were also observed in the ventricular CCS even at later developmental stages of 17.5 dpc (Fig. 4E) and in neonates at 4 days after birth (Fig. 4F). The AV cushion cells were weakly positive for the GFP signal, due to the thinness of the cytoplasm and resulting lower intensity of fluorescence (Fig. 4A), but their identity was confirmed by LacZ staining in *Mesp1-cre; R26R* embryos (Fig. 1C). Thus, we conclude unequivocally that the population of *Mesp1*-nonexpressing cells, which we identified along the ventricular septum, contributes to the CCS.

In the case of the SA or AV node regions of the CCS, the contribution of *Mesp1*-expressing and/or *Mesp1*-nonexpressing cells was not as clear from our present results using embryos at 13.5 dpc, because these typical node structures were not discernible. In contrast, we were able to determine that most of the cells in the venous valves, which are the proposed remnants of the embryonic SA ring in the fully developed heart (Rentschler et al., 2001), of the *Mesp1-cre:CAG-CAT-EGFP:CCS-lacZ* embryo were *Mesp1*-expressing (i.e., GFP–positive cells). This determination was revealed by the *Mesp1*-expressing/*CCS-lacZ*–positive yellow cells at 13.5 dpc (Fig. 4G). However, the developmental relationships between the venous valves and both the SA and AV nodes have not yet been determined.

Origin of *Mesp1*-Nonexpressing Cells

To determine the origin of the *Mesp1*-nonexpressing cells, we examined the *LacZ* expression profiles in more immature *Mesp1-cre:R26R* and *CCS-lacZ* embryos. As shown in Figure 5A, even at 7.75 dpc, at which stage the cardiac crescent can be observed, a few β-gal–negative cells were detectable in the *Mesp1-cre:R26R* embryo. The β-gal–negative cells were observed in the myocardium region more clearly at 8.5 dpc (Fig. 5B). In the *CCS-lacZ* embryo, although the heart region at 7.75 dpc was confirmed to be β-gal–negative (data not shown) as reported previously (Rentschler et al., 2001), patchy staining was observed mainly in the subendocardial myocardium region at 8.5 dpc (Fig. 5C). However, a direct relationship between the *Mesp1*-nonexpressing cells and the CCS cells is still not clear, although the neural crest cells, which are also identifiable as *Mesp1*-nonexpressing cells in our system, have not yet arrived in the heart at this stage and can be excluded (Jiang et al., 2000).

DISCUSSION

In this study, we have found using a Cre-*loxP* site-specific recombination system that the origin of the cardiac mesenchyme is subdivided according to the presence of *Mesp1* expression. We demonstrate that the regions occupied by *Mesp1*-nonexpressing cells correspond to two distinct populations of cells: one derived from the neural crest and the other one that contributes to the ventricular CCS.

Comparison of the Cell-Lineages of Neural Crest Cells and *Mesp1*-Nonexpressing cells

In our experiments with *Mesp1-cre:R26R* embryos, we have found that cells derived from the neural crest are negative but that mesodermal cells derived from *Mesp1*-expressing cells are positive, for β-gal activity. We have also confirmed that mammalian cardiac neural-crest cells are *Mesp1*-negative (Figs. 1, 2) and contribute to the mesenchyme in the OT cushions of the heart. These observations were made following neural crest cell lineage analyses using the *P0-cre:CAG-CAT-Z* strain (Fig. 2) and are consistent with previous results obtained using *Wnt1-cre:R26R* double transgenic mice (Jiang et al., 2000). The origin of the cells of the AV cushions was suggested to be mesodermal, because this region was occupied by *Mesp1*-expressing cells in our study (Fig. 1C). This result is consistent with the previous study of Kisanuki et al. (2001) using *Tie2-cre* mice that reported that the origin of the AV cushions is mainly of endocardial cell lineage. Thus, mesenchymal cells in the OT cushions are derived from mainly neural crest cells and those in the AV cushions are derived from endocardium.

Importantly, we observed a second population of *Mesp1*-nonexpressing cells, along the IVS (Fig. 1B). Because this region is not contributed by neural crest cells (Fig. 2B), we explored the possibility that these *Mesp1*-nonexpressing cells reside in the ventricular CCS. Before examining this possibility, we first confirmed that the failure to express β-gal in the *Mesp1-cre:R26R* embryos was not due to an artifact, such as down-regulation of *LacZ* expression during differentiation or mosaicism of Cre recombinase expression. To exclude the former possibility, we examined *CAG-cre:R26R* mice, in which Cre recombinase is ubiquitously expressed and all cells should be *LacZ*-positive. We did not subsequently observe any *LacZ*-negative cells in the heart, indicating that there had been no down-regulation of *LacZ* upon cell differentiation (Fig. 1D). To exclude possible mosaicism of Cre recombinase, we repeated our analysis in more than 20 embryos and observed very consistent results, although some clonal differences may exist. In addition, when we crossed the *Mesp1-cre* and *CCS-lacZ* strains and monitored *R26R*-dependent reporter gene expression, we did not observe patchy *LacZ*-negative cells in the ventricular wall. Thus, it appears unlikely that mosaicism of the *R26R* reporter could account for our results.

As for the contribution of the neural crest cells into the ventricular CCS, it was reported that neural crest-derived cells were observed in the vicinity of the CCS in the IVS at 14.5 dpc using the *Wnt1-cre:R26R* reporter system (Poelmann et al., 2004). Thus, the possibility cannot be rule out that the neural crest cells contribute to CCS in the IVS, although we could not detect any β-gal–positive cells in the IVS in our *P0-cre:CAG-CAT-Z* system. The discrepancy could be due to the difference in systems used for lineage analyses. The future studies using triple transgenic strategy (*Wnt1-cre:CAG-CAT-GFP:CCS-lacZ*) as used in our current study would be useful to discriminate the discrepancy.

Origins of the CCS

Using *Mesp1-cre:R26R* embryos, we identified a population of *Mesp1*-nonexpressing cells that were found to be distributed in the wall along the ventricular septum (Fig. 1B). The results of genetic crosses with the *CCS-lacZ* strain suggested that these *Mesp1*-nonexpressing cells contribute to the ventricular CCS (Fig. 3B). To confirm these findings, we generated triple transgenic *Mesp1-cre:CAG-CAT-EGFP:CCS-lacZ* mice. Double-staining for GFP and β-gal expression and/or additional DAPI staining in these mice confirmed that the *Mesp1*-nonexpressing cells contribute approximately 20% of the ventricular CCS (Fig. 4). Moreover, these populations of cells can be distinguished at a stage as early as stage 7.75 dpc at least (Fig. 5), whereas *Mesp1* is initially, albeit transiently, expressed at 6.5 dpc (Saga et al., 1996).

The pacemaking and conduction systems of the heart are composed of the SA node, AV node, AVB, the bundle branches, and the Purkinje fibers, each of which can be distinguished morphologically, functionally, and molecularly (Moorman and Christoffels, 2003). The origin of the nodal tissue is less clear than that of the ventricular CCS, although the primary myocardium is suggested to be a candidate (Moorman and Christoffels, 2003). Recently, it was suggested that some of the working myocardium could also differentiate into nodal tissues, even after birth (Pashmforoush et al., 2004). Although the developmental relationships between the venous valves and the nodes have not yet been fully elucidated, our data indicate that most of cells in the venous valves, which are proposed to be remnants of the embryonic SA ring (Rentschler et al., 2001), are derived from *Mesp1*-expressing cells (Fig. 4G). However, further detailed studies will be required to determine the precise cellular origin of the nodes and their relationships with the venous valves.

In the present analyses, we have focused on the cell-lineages of the ventricular CCS and shown them to be of

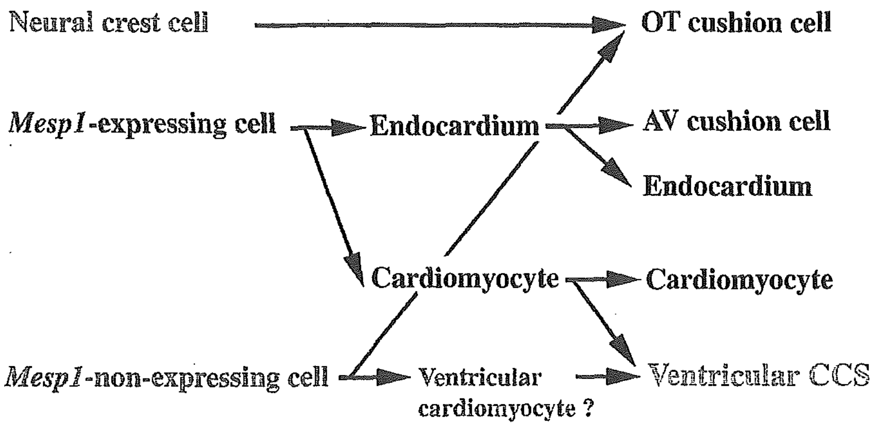

Fig. 6. Summary of the origin and cell-fate relationships of cardiac mesenchyme cell types. Each cardiac cell type is established by three distinct origins: neural crest cells, the mesodermal cells of *Mesp1*-expressing cells, and *Mesp1*-nonexpressing cells. It is noteworthy that both the *Mesp1*-expressing cells and the *Mesp1*-nonexpressing cells contribute to the ventricular CCS. In addition, the origins of the subset of the ventricular CCS that are contributed by the *Mesp1*-nonexpressing cells are distinguishable from that of the myocardium by the *Mesp1* expression profile. We speculate that the *Mesp1*-nonexpressing cardiomyocyte may be a candidate for the origin of the subset of the ventricular CCS. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

heterocellular origin. Two possibilities have emerged from both our analyses and previous reports concerning the origin of the ventricular CCS in mouse, occupied by *Mesp1*-nonexpressing cells. First, it is conceivable that the *Mesp1*-nonexpressing cells in the CCS are not derived from cardiomyocytes. Alternatively, these cells may represent cardiomyocytes, which simply do not express *Mesp1*. We favor this latter possibility. Lineage tracing experiments in chick have convincingly demonstrated that the ventricular CCS, including the Purkinje fibers, are derived from cardiomyocytes (reviewed in Mikawa, 1999; Pennisi et al., 2002). Moreover, experiments in the mouse also indicate that embryonic cardiomyocytes can be converted to a CCS-like phenotype in response to neuregulin-1, at least when assayed by up-regulation of the *CCS-lacZ* transgene (Rentschler et al., 2002). Nonetheless, additional analyses will be required to determine the basis for the molecular heterogeneity within the ventricular CCS and to determine whether there is associated functional diversity in this structure.

In conclusion, we have determined that *Mesp1*-nonexpressing cells contribute to the ventricular CCS in addition to the OT cushion. Furthermore, we indicate a possibility that a population of the cells that contribute to the ventricular CCS might be distinguished at an early stage of development. Unfortunately, it could not be clarified from our present experiments whether *Mesp1*-nonexpressing cells also contributed to the other regions of the CCS, such as the SA or AV nodes. A scheme summarizing the cell lineage relationships in the developing murine heart is shown in Figure 6. Our observation that the ventricular CCS includes both *Mesp1*-expressing and -nonexpressing cells is evidence of the heterogeneous nature of the ventricular CCS. The further identification of specific molecular markers for the mouse CCS, expressed at early embryonic stages, will undoubtedly enhance our understanding of the developmental biology of the CCS in the heart.

EXPERIMENTAL PROCEDURES

Lineage Analysis of *Mesp1*-Expressing Cells

The *Mesp1-cre* knockin mouse was constructed by introduction of a gene encoding Cre recombinase into the *Mesp1* locus, as previously described (Saga et al., 1999). The fidelity of expression was confirmed by in situ hybridization at E7.0 (data not shown). Genotyping was performed by polymerase chain reaction using a neo-specific primer NeoAL2: 5′-GGGGATGCGGTGGGCTCTATGGCTT-3′ and *Mesp1* primer MesP1-GR1: 5′-ATATGCCAAGTCATTGAGGTGAGCTTTC-3′. *Mesp1-cre* mice were crossed with either *CAG-CAT-Z* (Araki et al., 1995), *R26R* (Soriano, 1999), or *CAG-CAT-EGFP* (Kawamoto et al., 2000) reporter mice. *P0-cre* (Yamauchi et al., 1999) and *CCS-lacZ* mice (Rentschler et al., 2001) were also used for cell lineage analyses. Mice were maintained on a 7:00 AM to 7:00 PM light–dark cycle, with noon on the day of vaginal plug discovery defined as 0.5 dpc.

β-gal Staining, Immunostaining, and In Situ Hybridization

Embryos that had been fixed at 7.5–10.5 dpc were stained for the detection of β-galactosidase activity in wholemounts as described previously (Saga et al., 1992). The specimens were then dehydrated by means of a graded ethanol series, embedded in either paraffin wax or plastic resin (technovit 8100, Heraeus Kulzer, Inc.) and sectioned at a thickness of 4 μm. Hearts that had been isolated from embryos at later stages were subjected to β-gal staining after sectioning. Briefly, hearts were fixed in a solution of 2% paraformaldehyde, 0.05% glutaraldehyde, and 0.02% NP-40 in phosphate buffer (PBS) for 30 min on ice. The tissues were then sequentially soaked in a graded series of 10, 20, and 30% sucrose (w/v) in PBS while being gently agitated on a shaking platform, culminating in a 50:50 mix of 30% sucrose:OCT. Samples were frozen and stored at −80°C until sectioning at 8 μm thickness, and the sections were placed on gelatin-coated slides. Frozen sections of *Mesp1-cre:CAG-CAT-EGFP:CCS-lacZ* mouse hearts was stained with anti-lacZ and anti-GFP antibodies as follows: sections prepared were fixed with 4% paraformaldehyde for 3 min, treated with 10 μg/ml proteinase K and blocked in 3% skim milk for 30 min at room temperature (RT). Blocking solutions was replaced with rabbit anti–β-gal antibody (Cappel, ICN Pharmaceuticals, Inc., OH) at a dilution of 1:2,000 and with rat anti-GFP antibody (Nacalai Tesque, Kyoto, Japan) at a dilution 1:200 and incubated overnight at 4°C. After brief washes in PBS, the sections were incubated with Alexa 594–conjugated anti-rabbit followed by Alexa 488–conjugated anti-rat secondary antibodies at dilutions of 1:200

for 90 min at RT. These sections were then incubated with 0.1 μg/ml of DAPI (Sigma, St. Louis, MO) for 5 min to visualize nuclei.

ACKNOWLEDGMENTS

We thank the following researchers for providing mice: Dr. Jun-ichi Miyazaki (*CAG-cre*, *CAG-CAT-Z*, and *CAG-CAT-EGFP*), Dr. Kuniya Abe (*P0-cre*), Dr. Philippe Soriano (*R26R*), and Dr. Glenn I. Fishman (*CCS-lacZ*). We also thank Seiko Shinzawa, Mariko Ikumi, Eriko Ikeno, Chizuko Obata, Shinobu Watanabe, and Maho Endo for technical assistance.

REFERENCES

Araki K, Araki M, Miyazaki J, Vassalli P. 1995. Site specific recombination of a transgene in fertilized eggs by transient expression of Cre recombinase. Proc Natl Acad Sci U S A 92:160–164.

Cheng G, Litchenberg WH, Cole GJ, Mikawa T, Thompson RP, Gourdie RG. 1999. Development of the cardiac conduction system involves recruitment within a multipotent cardiomyogenic lineage. Development 126:5041–5049.

Coppen SR, Dupont E, Rothery S, Severs NJ. 1998. Connexin45 expression is preferentially associated with the ventricular conduction system in mouse and rat heart. Circ Res 82:232–243.

Coppen SR, Severs NJ, Gourdie RG. 1999. Connexin45 (alpha 6) expression delineates an extended conduction system in the embryonic and mature rodent heart. Dev Genet 24:82–90.

Delorme B, Dahl E, Jarry-Guichard T, Marics I, Briand JP, Willecke K, Gros D, Theveniau-Ruissy M. 1995. Developmental regulation of connexin 40 gene expression in mouse heart correlates with the differentiation of the conduction system. Dev Dyn 204:358–371.

Gassanov N, Er F, Zagidullin N, Hoppe UC. 2004. Endothelin induces differentiation of ANP-EGFP expressing embryonic stem cells towards a pacemaker phenotype. FASEB J 18:1710–1712.

Gourdie RG, Mima T, Thompson RP, Mikawa T. 1995. Terminal diversification of the myocyte lineage generates Purkinje fibers of the cardiac conduction system. Development 121:1423–1431.

Jiang X, Rowitch DH, Soriano P, McMahon AP, Sucov HM. 2000. Fate of the mammalian cardiac neural crest. Development 127:1607–1616.

Kawamoto S, Niwa H, Tashiro F, Sano S, Kondoh G, Takeda J, Tabayashi K, Miyazaki J. 2000. A novel reporter mouse strain that expresses enhanced green fluorescent protein upon Cre-mediated recombination. FEBS Lett 470: 263–268.

Kirby ML, Gale TF, Stewart DE. 1983. Neural crest cells contribute to normal aorticopulmonary septation. Science 220: 1059–1061.

Kisanuki YY, Hammer RE, Miyazaki J, Williams SC, Richardson JA, Yanagisawa M. 2001. Tie2-Cre transgenic mice: a new model for endothelial cell-lineage analysis in vivo. Dev Biol 230:230–242.

Kitajima S, Takagi A, Inoue T, Saga Y. 2000. MesP1 and MesP2 are essential for the development of cardiac mesoderm. Development 127:3215–3226.

Kupershmidt S, Yang T, Anderson ME, Wessels A, Niswender KD, Magnuson MA, Roden DM. 1999. Replacement by homologous recombination of the minK gene with lacZ reveals restriction of minK expression to the mouse cardiac conduction system. Circ Res 84:146–152.

Mikawa T. 1999. Cardiac lineages. In: Harvey RP, Rosenthal N, editors. Heart development. San Diego: Academic Press. p 19–33.

Moorman AFM, Christoffels VM. 2003. Cardiac chamber formation: development, genes, and evolution. Physiol Rev 83:1223–1267.

Moorman AFM, deJong F, Denyn MM, Lamers WH. 1998. Development of the cardiac conduction system. Circ Res 82: 629–644.

Myers DC, Fishman GI. 2004. Toward an understanding of the genetics of murine cardiac pacemaking and conduction system development. Anat Rec A Discov Mol Cell Evol Biol 280:1018–1021.

Nguyen-Tran VT, Kubalak SW, Minamisawa S, Fiset C, Wollert KC, Brown AB, Ruiz-Lozano P, Barrere-Lemaire S, Kondo R, Norman LW, et al. 2000. A novel genetic pathway for sudden cardiac death via defects in the transition between ventricular and conduction system cell lineages. Cell 102:671–682.

Pashmforoush M, Lu JT, Chen H, Amand TS, Kondo R, Pradervand S, Evans SM, Clark B, Feramisco JR, Giles W, Ho SY, Benson DW, Silberbach M, Shou W, Chien KR. 2004. Nkx2–5 pathways and congenital heart disease; loss of ventricular myocyte lineage specification leads to progressive cardiomyopathy and complete heart block. Cell 117:373–386.

Pennisi DJ, Rentschler S, Gourdie RG, Fishman GI, Mikawa T. 2002. Induction and patterning of the cardiac conduction system. Int J Dev Biol 46:765–775.

Poelmann RE, Jongbloed MR, Molin DG, Fekkes ML, Wang Z, Fishman GI, Doetschman T, Azhar M, Gittenberger-de Groot AC. 2004. The neural crest is contiguous with the cardiac conduction system in the mouse embryo: a role in induction? Anat Embryol (Berl) 208:389–393.

Rentschler S, Vaidya DM, Tamaddon H, Degenhardt K, Sassoon D, Morley GE, Jalife J, Fishman GI. 2001. Visualization and functional characterization of the developing murine cardiac conduction system. Development 128:1785–1792.

Rentschler S, Zander J, Meyers K, France D, Levine R, Porter G, Rivkees SA, Morley GE, Fishman GI. 2002. Neuregulin-1 promotes formation of the murine cardiac conduction system. Proc Natl Acad Sci U S A 99:10464–10469.

Saga Y. 1998. Genetic rescue of segmentation defect in MesP2-deficient mice by MesP1 gene replacement. Mech Dev 75: 53–66.

Saga Y, Yagi T, Ikawa Y, Sakakura T, Aizawa S. 1992. Mice develop normally without tenascin. Genes Dev 6:1821–1831.

Saga Y, Hata N, Kobayashi S, Magnuson T, Seldin M, Taketo MM. 1996. MesP1: a novel basic helix-loop-helix protein expressed in the nascent mesodermal cells during mouse gastrulation. Development 122:2769–2778.

Saga Y, Hata N, Koseki H, Taketo MM. 1997. Mesp2: a novel mouse gene expressed in the presegmented mesoderm and essential for segmentation initiation. Genes Dev 11:1827–1839.

Saga Y, Miyagawa-Tomita S, Takagi A, Kitajima S, Miyazaki J, Inoue T. 1999. MesP1 is expressed in the heart precursor cells and required for the formation of a single heart tube. Development 126: 3437–3447.

Saga Y, Kitajima S, Miyagawa-Tomita S. 2000. *Mesp1* expression is the earliest sign of cardiovascular development. Trends Cardiovasc Med 10:345–352.

Sakai K, Miyazaki J. 1997. A transgenic mouse line that retains Cre recombinase activity in mature oocytes irrespective of the cre transgene transmission. Biochem Biophys Res Commun 237:318–324.

Soriano P. 1999. Generalized lacZ expression with the ROSA26 Cre reporter strain. Nat Genet 21:70–71.

Waldo K, Miyagawa-Tomita S, Kumiski D, Kirby ML. 1998. Cardiac neural crest cells provide new insight into septation of the cardiac outflow tract: aortic sac to ventricular septal closure. Dev Biol 196: 129–144.

Yamauchi Y, Abe K, Mantani A, Hitoshi Y, Suzuki M, Osuzu F, Kuratani S, Yamamura K. 1999. A novel transgenic technique that allows specific marking of the neural crest cell lineage in mice. Dev Biol 212:191–203.

Mass Distributed Clustering: A New Algorithm for Repeated Measurements in Gene Expression Data

Shinya Matsumoto[1,*,†] **Ken-ichi Aisaki**[2,*] **Jun Kanno**[2,*,‡]
shinya.matsumoto@ncr.com aisaki@nihs.go.jp kanno@nihs.go.jp

[1] Teradata Division, NCR Japan, Ltd. 2-4-1 Shiba-koen, Minato-ku Tokyo 105-0011, Japan
[2] Cellular & Molecular Toxicology, Biological Safety Research Center, National Institutes of Health Sciences, 1-18-1 Kamiyoga, Setagaya-ku Tokyo 158-8501, Japan

Abstract

The availability of whole-genome sequence data and high-throughput techniques such as DNA microarray enable researchers to monitor the alteration of gene expression by a certain organ or tissue in a comprehensive manner. The quantity of gene expression data can be greater than 30,000 genes per one measurement, making data clustering methods for analysis essential. Biologists usually design experimental protocols so that statistical significance can be evaluated; often, they conduct experiments in triplicate to generate a mean and standard deviation. Existing clustering methods usually use these mean or median values, rather than the original data, and take significance into account by omitting data showing large standard deviations, which eliminates potentially useful information. We propose a clustering method that uses each of the triplicate data sets as a probability distribution function instead of pooling data points into a median or mean. This method permits truly unsupervised clustering of the data from DNA microarrays.

Keywords: data mining, bioinformatics, gene expression data, microarray, repeated measurements, clustering algorithm

1 Introduction

1.1 Motivation

When large-scale gene expression profiles became available, biologists usually normalized the data to overt biological events, such as monitorable phenotypes. By doing this, biologically important expression data could be selected by linkage analysis to a particular biological events and used for further analysis. This type of analysis tends to be limited to genes that encode the final phases of a gene cascade or signaling system that directly reflects an emergence of phenotype, and hence shows high expression values.

The advent of microarray and other high-throughput technologies has removed such limitations, allowing whole-genome analysis that includes the initial phases of the cascade, where phenotypes are not clear and signal intensity is usually low. These technologies generate huge quantities of data, placing great demands on data analysis. To accommodate this demand, the Division of Cellular and Molecular Toxicology of National Institute of Health Sciences (NIHS), Japan, has developed the Percellome System [2], which generates absolute mRNA-quantity data as the copy number per cell from the microarray system and quantitative PCR. This system essentially enables utilization of all

[*] These authors contributed equally to this work.

[†] To whom correspondence about mathematical issues should be addressed: E-mail: shinya.matsumoto@ncr.com

[‡] To whom correspondence about biological issues including Percellome system should be addressed: E-mail: kanno@nihs.go.jp

Table 1: Sample data.

	Condition 1			Condition 2		
	1st Exp.	2nd Exp.	3rd Exp.	1st Exp.	2nd Exp.	3rd Exp.
Gene 1	0.9682	0.9924	1.0394	-0.1277	-0.0842	0.2125
Gene 2	1.3656	1.4547	1.3798	-0.2026	-0.2539	0.4596
Gene 3	-0.0109	-0.0619	0.0738	0.9116	0.9532	1.1352
Gene 4	-0.1315	-0.0222	0.1540	1.3569	1.2596	1.5835
Gene 5	-1.1195	-0.9738	-0.9067	0.0605	0.0946	-0.1543
Gene 6	-1.4476	-1.2152	-1.5372	0.0088	0.0508	-0.0587
Gene 7	-0.0070	0.0697	-0.0623	-0.8928	-1.0297	-1.0775
Gene 8	-0.1236	-0.2152	0.3397	-1.3814	-1.3456	-1.4730
Gene 9	1.2004	0.0041	1.0455	-0.2224	0.5194	0.4527
Gene 10	0.1282	0.4077	0.2144	1.1292	0.4488	0.6720
Gene 11	-1.2166	0.2551	0.2115	1.3180	0.7994	0.1325
Gene 12	-0.5777	-0.7242	-0.9481	-0.0263	0.1748	0.6009
Gene 13	-0.2747	-0.3692	-1.6061	-0.8657	-0.0627	0.1783
Gene 14	0.6377	0.1786	-1.5665	-1.2766	-0.7981	-0.1753
Gene 15	-1.1518	0.9327	0.9700	-1.2910	-0.8788	-0.0802
Gene 16	0.0885	1.7689	0.3925	0.0623	0.8591	-1.6715
Gene 17	1.3529	1.8681	-1.7204	1.8635	0.2069	-0.5710
Gene 18	0.7227	0.0423	0.7346	-0.8883	-0.1600	-0.4517
Gene 19	-1.4129	-0.2668	0.1797	1.0153	2.5328	-2.0493
Gene 20	-0.2813	-0.6737	-0.5450	0.4369	-1.0448	-0.8920
Gene 21	-0.2935	2.4749	3.2186	-1.8833	1.6711	0.2152
Gene 22	-0.3168	0.3275	-5.4107	2.0824	0.9931	-3.0695
Gene 23	0.0022	0.1320	-0.1333	2.6916	-0.3704	3.0789
Gene 24	-3.1931	-0.6846	3.8781	-1.8794	-2.8393	-0.6813

of the gene expression data for the clustering analysis. The basis of the clustering strategy for this all-gene data is a phenotype-independent analysis, meaning that there are no auxiliary data that can be used for clustering. We have designed a pure, unsupervised clustering system that can handle low-intensity data along with its variance. It is postulated that low-intensity data may contain relatively larger amounts of measurement error than high-intensity data.

An example of the data collected for this study is shown in Table 1. Two experiments were performed in triplicate (i.e., each experiment was performed on three mice). An average result for each replicated experiment can be calculated, but this eliminates information about any deviations. Alternatively, the data from the three replicates can be treated as a probability distribution and handled by a parametric approach. Applying this approach to gene expression data, we were able to develop our unsupervised clustering algorithm, mass distributed clustering (MADIC).

1.2 Related Works

Most clustering algorithms ignore measurement errors. However, measurement errors occur in the real world, especially in gene expression analysis. NIHS has established a measurement method that can analyze all genes, including low-intensity genes that contain substantial measurement errors [2]. Some clustering algorithms, such as the one presented by Kumar et al., can handle data with errors [3]. Yeung et al. demonstrated clustering algorithms in which deviations from repeated measurements can be evaluated [6]. Using a clustering algorithm with SD- or CV-distance, their approach was an improvement over the traditional simple average method, which does not evaluate deviations.

Many clustering algorithms have been proposed for gene expression analysis [4, 5], but existing algorithms cannot use whole genes, stable and unstable genes. Such methods are useful when clustering stable objects, but we wanted to devise a method to cluster both stable and unstable genes. Our proposed algorithm is based on an extension of density-based clustering, DBSCAN [1].

1.3 Purpose of Research

The purpose of this research was to develop a clustering algorithm that would handle triplicate gene expression data without losing information about deviation.

1.4 Outline of Article

In Section 2, we define how we have extended density-based clustering and how our proposed algorithm differs from those used in conventional clustering. In Section 3, we provide the details of our clustering algorithm. In Section 4, we present results of experiments with synthetic and real gene data. Finally, in Section 5 we offer conclusions.

2 Definitions

2.1 Notations

(1) Data set. $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_n\}$: Data set for clustering. $\mathbf{o}_i$ is a probability distribution function (PDF) in d-dimensional Euclidean space. We sometimes denote this by the objects $\mathbf{o}, \mathbf{p}, \mathbf{q} \in \mathbf{O}$. $\mathbf{o}_i := p_{\sigma_i}(x_i) : \mathbf{R}^d \rightarrow \mathbf{R}^+$: PDF. $\mathbf{x}_i \in \mathbf{R}^d$ is a point of d-dimensional Euclidean space. σ_i is the parameter of the PDF. We also represent the PDF in another way. We can use this notation if the PDF has no special direction for this integration. $p_{\mathbf{x}_i\sigma_i}(\mathbf{r}) : \mathbf{R}^+ \rightarrow \mathbf{R}^+$: PDF. $\mathbf{r}$ is the distance from $\mathbf{x}_i$.

(2) Observation. $\mathbf{y}(i, j, k)$: kth observed value for jth dimension for $\mathbf{o}_i$.

(3) Distance. Distance is denoted by $\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j)$ in usual Euclidean space. We define distance between objects as $\mathrm{dist}(\mathbf{o}_i, \mathbf{o}_j) = \mathrm{dist}(p_{\sigma_i}(\mathbf{x}_i), p_{\sigma_j}(\mathbf{x}_j)) = \mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j)$.

(4) ε. We use ε for the threshold about distance.

(5) θ_m. We use θ_m for threshold about mass.

(6) Mass function. We defined the mass function as:

$$\frac{\partial}{\partial \mathbf{r}} \mathbf{m}_\sigma(\mathbf{r}) = \mathbf{p}_{\mathbf{x},\sigma}(\mathbf{r}).$$

We sometimes denote the mass function and PDF with an object index such as $\mathbf{m}_{\mathbf{o}_2}$, which means $\mathbf{m}_{\sigma_2}$. The mass function has to have the following properties.

a) $\mathbf{m}_\sigma(0) = 0$: additional definition.

b) Increasing function: the definition is the differentiation form and the right side is equal to or greater than zero.

c) If $\mathbf{m}_{\sigma_1}(\mathbf{r}) > \mathbf{m}_{\sigma_2}(\mathbf{r})$ for some $\mathbf{r} > 0$, then the inequality is true for any positive number.

d) $\mathbf{m}_\sigma(\infty) = 1$: convenient for giving algorithm parameters.

Well-known probability distributions, such as the chi-square function, have these properties. The function $\mathbf{p}$ is the PDF. The mass function $\mathbf{m}$ is the cumulative PDF.

2.2 The Expansion of Density-Based Clustering

We expand and redefine definitions used in traditional density-based clustering as follows:

(1) ε-neighborhood. *ε-neighborhood* of an object $\mathbf{p}$, denoted by $\mathbf{N}_\varepsilon(\mathbf{p})$, is defined by $\mathbf{N}_\varepsilon(\mathbf{p}) = \{\mathbf{q} \in \mathbf{O} : \mathrm{dist}(\mathbf{p}, \mathbf{q}) < \varepsilon\}$. This is the subset of the whole data set that has a distance less than ε.

(2) ε-neighborhood mass. ε-neighborhood mass of an object $\mathbf{p}$, denoted by $\mathbf{M}_\varepsilon(\mathbf{p})$, is defined by $\mathbf{M}_\varepsilon(\mathbf{p}) = \sum_{\mathbf{q} \in \mathbf{N}_\varepsilon(\mathbf{p})} \mathbf{m}_\mathbf{q}(\varepsilon - \mathrm{dist}(\mathbf{p}, \mathbf{q}))$.

Figure 1 shows the concept of ε-neighborhood mass in one dimension. This example shows ε-neighborhood mass of the center object. The center object is summarized within the radius ε, because $\text{dist}(\mathbf{p},\mathbf{p}) = 0$. The mass of center object is represented by the horizontally filled area, including its left-most expansion into the vertically filled area (crosshatched). There are two objects within ε except for own object, center object. These objects are summarized within the radius ε-dist $(\mathbf{p},\mathbf{q})$. The mass of left object is represented by the vertically filled area, including its overlap with the horizontally filled area (crosshatched). The mass of right object is diagonally filled area. Summing the three masses, $\mathbf{m_q}(\varepsilon\text{-dist}(\mathbf{p},\mathbf{q}))$, gives $\mathbf{M}_\varepsilon(\mathbf{p})$. The Crosshatched in this example is double counted.

ε-neighborhood mass is the expansion of ε-neighborhood. It supposes an infinite limit. If the mass is concentrated in the center, ε-neighborhood mass equals the number of objects in ε-neighborhood.

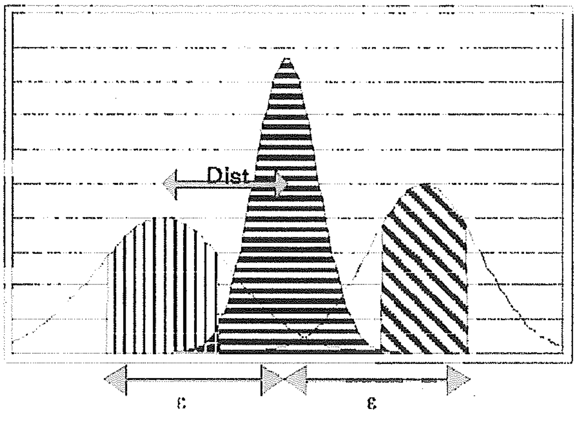

Figure 1: ε-neighborhood mass.

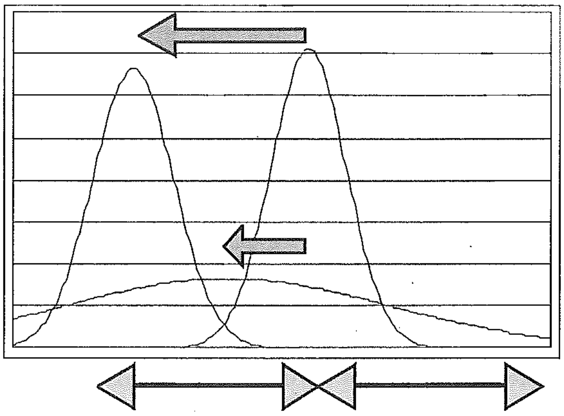

Figure 2: Directly density reachable.

(3) Directly density reachable. An object $\mathbf{p}$ is directly density reachable from an object $\mathbf{q}$ wrt. ε and $\mathbf{q_m}$ if:

a) $\mathbf{p} \in \mathbf{N}_\varepsilon(\mathbf{q}) \subset \mathbf{O}$.

b) $\mathbf{N}_\varepsilon(\mathbf{q}) > \theta_m$ (core condition 1).

c) $\mathbf{m}_q(\varepsilon) > \mathbf{m}_p(\varepsilon)$ (core condition 2).

Core condition 1 is the natural expansion of the core condition in DBSCAN. Core condition 2 shows the direction of the error rate in the experiment. Figure 2 shows the concept of core condition 2. This condition represents flow from a concentrated object to a distributed object, or from a high-density object to a low-density object.

(4) Density reachable. An object $\mathbf{p}$ is density reachable from an object $\mathbf{q}$ wrt. ε and θ_m if there is a chain of objects $\mathbf{p}_1, \cdots, \mathbf{p}_n, \mathbf{p}_1 = \mathbf{p}, \mathbf{p}_n = \mathbf{q}$ such that $\mathbf{p}_{i+1}$ is directly density reachable from $\mathbf{p}_i$. Figure 3 shows the concept of density reachable. This definition is the same as the DBSCAN definition.

(5) Density connected. Density connectivity is a symmetric relation. An object $\mathbf{p}$ is density connected to an object $\mathbf{q}$ wrt. ε and θ_m if there is a chain of objects $\{\mathbf{o}_1, \mathbf{p}_1, \mathbf{q}_1, \mathbf{o}_2, \mathbf{p}_2, \mathbf{q}_2, \cdots, \mathbf{p}_{m-1}, \mathbf{q}_{m-1}, \mathbf{o}_m\}$ such that:

a) Object $\mathbf{p}$ is density reachable from object $\mathbf{o}_1$.

b) Object $\mathbf{q}$ is density reachable from object $\mathbf{o}_m$.

c) Object $\mathbf{p}_i$ is density reachable from object $\mathbf{o}_i$.

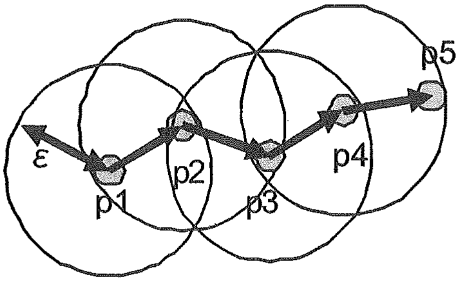

Figure 3: Density reachable.

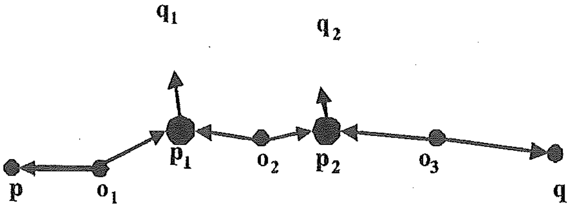

Figure 4: Density connected.

d) Object $\mathbf{p}_i$ is density reachable from object $\mathbf{o}_{i+1}$.

e) Object $\mathbf{q}_i$ is density reachable from object $\mathbf{p}_1$.

Because density reachable is defined as flow from a stable object to an unstable object, we cannot define density connected using one object as is done in DBSCAN, so instead we use the chain of objects. Condition 5 shows that each object, $\mathbf{p}_i$, is in the core condition. Figure 4 shows an example of density connected. The arrow is the flow of density reachable.

(6) Cluster. A cluster $\mathbf{C}$ wrt. ε and θ_m is a non-empty subset of $\mathbf{O}$ satisfying the following conditions:

a) If any $\mathbf{p} \in \mathbf{O}$ satisfies the core conditions, then $\mathbf{p}$ is a member of some cluster.

b) For any $\mathbf{p}, \mathbf{q} \in \mathbf{O}$: if $\mathbf{p}$ is a member of $\mathbf{C}$ and $\mathbf{q}$ is density connected from $\mathbf{p}$ wrt. ε and θ_m, then $\mathbf{q}$ is a member of $\mathbf{C}$ (maximality).

c) For any $\mathbf{p}, \mathbf{q} \in \mathbf{C}$: $\mathbf{p}$ is density connected to $\mathbf{q}$ wrt. ε and θ_m (connectivity).

A cluster contains the objects that do not satisfy the core condition. Such an object is called a *border*, and a border object may belong to multiple clusters.

2.3 Imitative Hierarchical Tree Structure

2.3.1 Lemmas

According to the preceding definitions, the following lemmas are true.

Lemma 1. If an object $\mathbf{p}$ is a core object wrt. ε_1 and θ_m, object $\mathbf{p}$ is a core object wrt. $\varepsilon_2 > \varepsilon_1$ and θ_m.

Proof. An object $\mathbf{p}$ is a core object wrt. ε_1 and $\mathbf{q}_m$. This means the following:

(1) $\mathbf{M}_{\varepsilon_1}(\mathbf{p}) > \theta_m$ (core condition 1).

(2) $\exists \mathbf{q} \in \mathbf{N}_{\varepsilon_1}(\mathbf{p}) \subset \mathbf{O}$ s.t. $\mathbf{m}_p(\varepsilon_1) > \mathbf{m}_q(\varepsilon_1)$ (core condition 2).

Because $\mathbf{M}_\varepsilon(\mathbf{p})$ is a strictly increasing function for ε, $\mathbf{M}_{\varepsilon_2}(\mathbf{p}) \geq \mathbf{M}_{\varepsilon_1}(\mathbf{p}) > \theta_m$ for $\varepsilon_2 > \varepsilon_1$. According to the mass function property (c), $\mathbf{m}_p(\varepsilon_1) > \mathbf{m}_q(\varepsilon_1) \Longrightarrow \mathbf{m}_p(\varepsilon_2) > \mathbf{m}_q(\varepsilon_2)$. And, according to the epsilon neighborhood, $\mathbf{q} \in \mathbf{N}_{\varepsilon_1}(\mathbf{p}) \subset \mathbf{N}_{\varepsilon_2}(\mathbf{p}) \subset \mathbf{O}$. So, $\exists \mathbf{q} \in \mathbf{N}_{\varepsilon_2}(\mathbf{p}) \subset \mathbf{O}$ s.t. $\mathbf{mp}(\varepsilon_2) > \mathbf{mq}(\varepsilon_2)$.

Lemma 2. If a subset $\mathbf{C}$ is a cluster wrt. ε_1 and θ_m, there is a cluster that contains $\mathbf{C}$ wrt. $\varepsilon_2 > \varepsilon_1$ and θ_m.

Proof. Suppose a subset $\mathbf{C}$ is a cluster wrt. ε_1 and θ_m. According to the connectivity condition, any objects $\mathbf{p}, \mathbf{q} \in \mathbf{C}$ are density connected. There exists a chain of objects which consists of directly density reachable or density reachable objects. These definitions are valid for $\varepsilon_2 > \varepsilon_1$, if satisfied for

ε_1. So, **p** and **q** are density connected for ε_2. According to the maximality condition, **p** and **q** are members of the same cluster.

Lemma 3. If an object **p** is a core object wrt. ε and θ_{m_1}, object **p** is a core object wrt. ε and $\theta_{m_2} < \theta_{m_2}$.

Proof is the same as Lemma 1.

Lemma 4. If a subset **C** is a cluster wrt. ε and θ_{m_1}, there is a cluster that contains **C** wrt. ε and $\theta_{m_2} < \theta_{m_1}$.

Proof is the same as Lemma 2.

2.3.2 Tree Structure

By proceeding with Lemmas 1, 2, 3 and 4, we can build a hierarchical tree structure if we use the appropriate thresholds and cluster the data. We call this structure a *imitative hierarchical tree structure* to distinguish it from hierarchical clustering.

For example, a sequence of thresholds is $\{\{\varepsilon_1, \theta_{m_1}\}, \{\varepsilon_2, \theta_{m_2}\}, \cdots, \{\varepsilon_n, \theta_{m_n}\}\}$, and a sequence of clusters $\{\{\mathbf{C}_{11}, \mathbf{C}_{12}, \cdots\}, \{\mathbf{C}_{21}, \mathbf{C}_{22}, \cdots\}, \cdots, \{\mathbf{C}_{n1}, \mathbf{C}_{n2}, \cdots\}\}$ correspond to the thresholds. For any cluster $\mathbf{C}_{ij}$ and $k < i$, there exists a cluster $\mathbf{C}_{km}$ such that $\mathbf{C}_{km}$ includes $\mathbf{C}_{ij}$. Figure 5 shows a tree structure. Each rectangle indicates cluster.

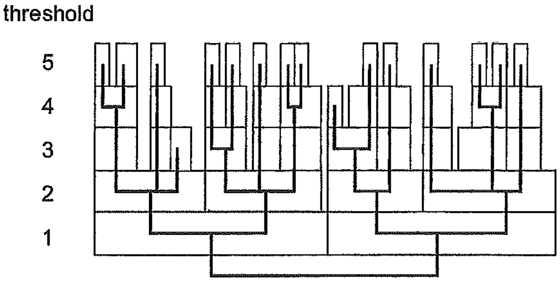

Figure 5: Imitative hierarchical thee stucture.

3 Algorithm

3.1 Our Solution

Our proposed algorithm is based on the following ideas:

(1) Consider the deviation of experimental data to be a mass distribution.

(2) Expand density-based clustering for the mass distribution.

(3) Generate the imitative hierarchical clustering tree to adapt the local density.

The deviation in data from identical replicate experiments can be represented as a PDF, and we identify the probability distribution with the mass distribution. By expanding density-based clustering, we created an algorithm to calculate the mass distribution as density. The density of DBSCAN is an integer number that represents the number of objects; in our algorithm, density is a real number. In using our algorithm, unstable genes should not be the core of a cluster, but in sparse regions the criteria of stableness should be loose. Our algorithm clusters for multiple thresholds and generates the imitative hierarchical tree, then chooses the appropriate clusters to adapt the local density.

3.2 Probability Distribution Function

We used the *gamma distribution function* as our PDF because cumulative gamma distributions have curves that are shaped like those of chi-square functions. A gamma distribution is a one-dimensional function that gives the distance from the center of an object:

$$p_{\alpha,\beta}(r) = \frac{1}{\beta^\alpha \Gamma(\alpha)} r^{\alpha-1} e^{-r/\beta}.$$

The cumulative gamma distribution is:

$$D_{\alpha,\beta}(r_0) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_{r_0}^{\infty} x^{\alpha-1} e^{-r/\beta} \mathrm{d}r.$$

We defined the two parameters for a gamma distribution as follows:

$$\alpha = \frac{d}{2}, \ \ \beta = \frac{2\sigma^2}{\alpha}.$$

A gamma function has the following properties:

(1) It is possible to calculate the integral function if alpha is a positive integer; it is called an *incomplete* gamma function.

(2) It is most dense around the center and least dense far from the center.

(3) The same deviation must be present in all directions. This condition can be difficult to meet for many domains, but it works for gene expression data because they have the same scale.

After normalizing our data, we defined the mass function as follows:

$$\mathbf{m}_\sigma(\mathbf{r}) = 1 - D_{\alpha,\beta}(\mathbf{r}^2) = 1 - D_{d/2, 2\sigma^2/d}(\mathbf{r}^2).$$

3.3 Algorithm on Threshold

It is difficult to determine what the threshold should be. An observation error changes the value of gene expression. Because of this, we do not cluster with a single threshold, but make imitative hierarchical clusters by changing threshold values. In this case, we give a threshold at appropriate intervals to perceive to a bigger change, than to perceive a change of the cluster constitution by changing of the delicate value of a threshold.

If there is a pure binary tree structure, the number of relationship within clusters is a power of 2. Figure 6 shows the relationship between tree structure and the relationships within clusters. The threshold marked by a double line indicates the smallest clusters; each cluster contains two objects and four relationships between objects. The threshold marked by a triple line indicates the next-level clusters; each cluster contains four objects and sixteen relations.

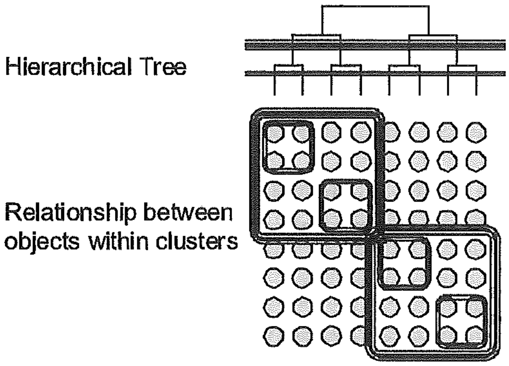

Figure 6: Hierarchical tree and relationship between objects within clusters.

We decided to use a rank of the distance between objects. We assigned the ranks using the following formula:

$$\text{Rank} := 10^{i/L} \ \ (i = 1, 2, \cdots)$$

Where ε_1 is defined as the 1st nearest distance, $10^{1/L}$, ε_2 is defined as the 2nd nearest distance, $10^{2/L}$ and so on.

3.4 Representation

After clustering for the threshold, each object is classified as core, as border, or as not belonging to any cluster. According to the Lemmas, when classified with a core object with a certain threshold, an object is always classified as a core object with a bigger threshold than it. It is thus possible to express core objects with a hierarchical tree structure.

Density-based clustering can find arbitrarily shaped clusters, but in gene expression analysis we want to find the clusters that have similar sizes. In the hierarchical cluster, we can find the appropriate cluster that satisfies the size condition.

3.4.1 Appropriate Cluster

For each threshold, we calculate the *diameters* of the clusters. Diameter is defined as the maximum distance between the core objects that belong to the cluster. We define the appropriate cluster as having a diameter less than the threshold and having the maximum diameter for the object.

3.4.2 Classification

We call the core objects of the appropriate cluster *rigorous* objects. Core objects that do not belong to an appropriate cluster but which are objects for the loosest threshold are called *shell* objects if they are direct-density-reachable from some rigorous objects, or *adhesive* objects if they are not direct-density-reachable from any rigorous object. The shell objects belong to the cluster that has the nearest rigorous object. There are some objects that are not core objects for the loosest threshold, and we group these into two types. First, the objects that satisfy core condition 1 and do not satisfy core condition 2 are called *unique* objects. These objects satisfy the mass threshold by themselves but they are far from other objects. The remaining objects are classified as *unstable.* All objects are classified into one of these four groups.

4 Experiments

4.1 Experiment with 2-Dimensional Synthetic Data

4.1.1 Data

The data in the 2-dimensional experiment consisted of 24 objects: 8 objects belonged to the clusters, the others were unstable objects. For each object, 100 points were generated, for a total of 2,400 points. Figure 7 illustrates the data. The four clusters and large amount noise are apparent.

Figure 8 shows the data from three points for each object. The clusters here are much more difficult to see. The difference between Figures 7 and 8 is due to the different number of observations.

Figure 10 shows the average value for each object. The black objects have small errors, whereas the gray objects have large errors. As in Figure 7, four clusters are visible (they are the eight black objects). The 16 gray objects represent background noise. Our algorithm works like Figure 9.

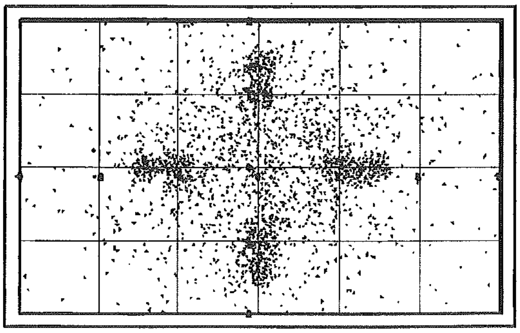

Figure 7: 100 points/object.

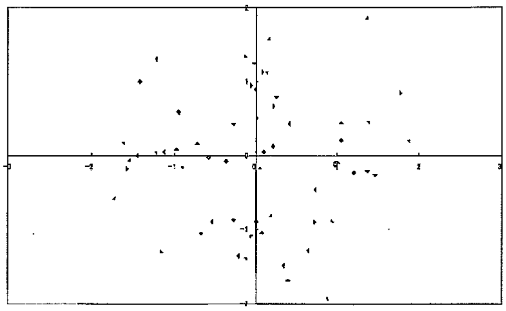

Figure 8: 3 points/object.

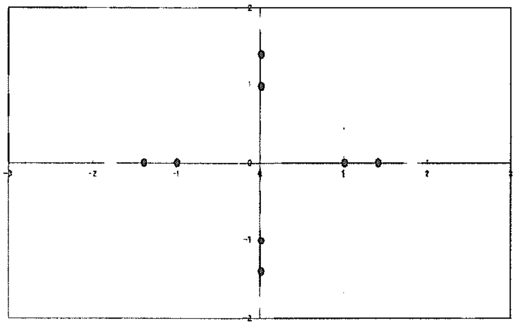

Figure 9: Average points.

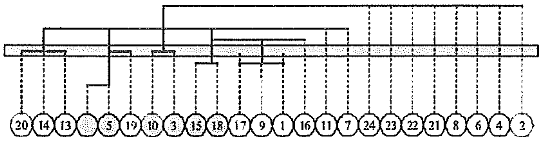

Figure 10: Hierarchical tree results from density-based clustering.

4.1.2 Results of Density-Based Clustering

We attempted to cluster with density-based clustering. In this two-dimensional test, we removed the normalization using the z-score. Figure 10 shows the results from using the density-based clustering algorithm with an imitative hierarchical tree.

Figure 11 is the 2 dimensional plot, which represents the clusters with the gray-labeled threshold in Figure 10. The black dots represent the core objects and the dotted ellipses show each cluster's core objects. The clusters in Figures 7 and 11 seem to be unrelated.

4.1.3 Result of MADIC

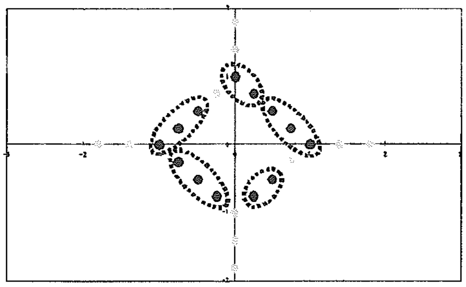

Figure 11: Results of density-based clustering.

Figure 12 shows the results of our algorithm. This figure shows the hierarchical tree and appropriate clusters. Our algorithm found five clusters.

Figure 13 shows the classification results. The rigorous objects are black and an ellipse surrounds each cluster. Four of five clusters are the same as those seen in Figure 7. The fifth cluster, in the bottom right quadrant, contains object No. 18. This object has less deviation data than the producing rules. This finding illustrates that errors sometimes affect results. But, it appeared in the loosest threshold. We should use the results with the threshold that appeared. In gene expression analysis, we do analyze the clusters that appeared in the severe threshold.

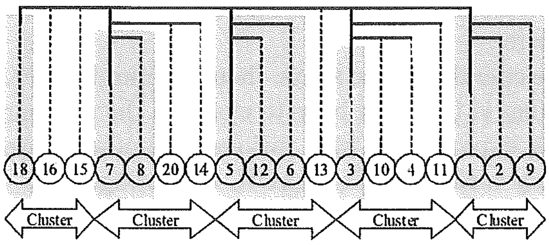

Figure 12: Results of our algorithm.

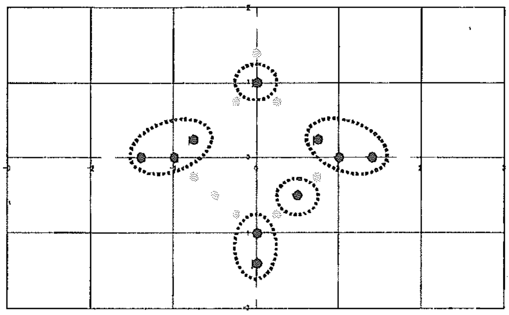

Figure 13: Classification results.

4.1.4 Comparison

Figure 14 shows the analysis flow. The generation rules exist but are always hidden; the objective of data mining is to discover these generation rules. If there are many observation points, we can discover the clusters, as seen in Figure 7. However, we sometimes obtain a limited number of observations.

We consider Figure 13 better than Figure 11. The proposed algorithm solves the new clustering problem. However, we cannot mathematically compare our algorithm and the existing algorithm.

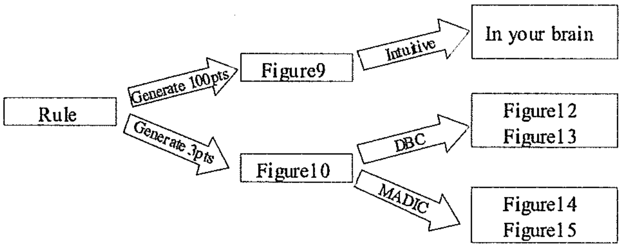

Figure 14: Analysis flow diagram.

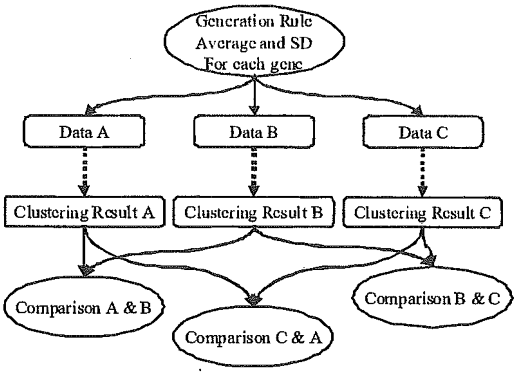

Figure 15: Synthetic data generation.

4.2 Experiment with 16-Dimensional Synthetic Data

4.2.1 Data

We used three data sets, each produced using the same rule. Each data set has 10,000 objects and 16 dimensions i.e. experimental conditions, and three data for each condition. Figure 15 shows the data generation, clustering, and comparison. We synthesized the average surfaces and then added to them random numbers. These data sets have the same characteristics, and therefore should generate the same clusters. However, the addition of random numbers generates the differences.

4.2.2 Evaluation Method

We added a big random number to simulate what we would see with real gene expression data. It is difficult to put them into the same cluster. We evaluated the number calculated by the following equation:

$$\text{Index} = \frac{\sum_{i,j=1}(\sharp(C_iC_j))^2}{\sqrt{\sum_{i=1}(\sharp(C_i))^2}\sqrt{\sum_{j=1}(\sharp(C_j))^2}}.$$

This equation shows the sum of squares of the number of intersections of both clusters, divided by the square roots of the sum of squares of the number of clusters.

This number is bounded from 0 to 1. If two clusters are completely the same, then this number is 1. If whole genes are grouped into one cluster, then this number is 1. So, we have to evaluate the number according to the number of clusters. We do the clustering experiments with various parameters. We gave the various numbers of clusters for k-means clustering various θ_m for MADIC clustering.

4.2.3 Results

We performed k-means clustering and MADIC clustering with various parameters (Figure 16). The indexes of the MADIC method are higher than those for k-means clustering. This result indicates that MADIC clustering is less affected by the random numbers than is k-means clustering.

4.3 Real Data

4.3.1 Data

NIHS performed experiments to determine how gene expression varied with exposure to four doses of thalidomide (vehicle only, low-dose, mid-dose, and high-dose) and four time points (2, 4, 8, and 24 hours later), i.e., 16 condition points with MOE430A of Affymetrix. Three mice were measured for each condition, creating triplicate data. Thus, 48 chips were used for the experiment.

Thalidomide is a drug for a sleeping aid and a treatment for morning sickness. It was subsequently found to be teratogenic, particularly during the first 25 to 50 days of pregnancy, most visibly causing amelia or phocomelia. The normalization was done by Percellome System for these measurement results. k-means clusterings were done for 16-dimensional data, averages for each condition with k = 80, 90, 100, compared with the MADIC result. It is difficult to predetermine the number of partitions, which is a very important parameter in k-means. AIC with EM method gave a partition number of one, which is obviously unacceptable; normally the number of clusters can only be determined from a biological viewpoint. In this study, partition numbers for k-means were the number of clusters given by MADIC.

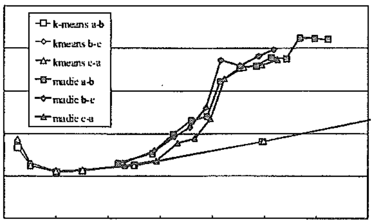

Figure 16: Results of 16-dimensional synthetic data.

4.3.2 Results

MADIC found 138 rigorous probes and 87 clusters. Moreover, 122 unique probes were found. These probes in clusters were summed up by the keywords in the Gene Ontology Biological Process. Ratios were calculated that had the same keywords in each cluster. Figure 17 shows the distributions of number of clusters for highest ratio of keywords in the Biological Process. MADIC could identify the clusters which have ratio more than 18%. This means that MADIC could find clusters which are assumed having biological meaning. Table 2 shows the average of the highest ratios of the same keyword. MADIC was higher than the k-means values. It is thought that this reflects division into the cluster that belongs to same key word in the Biological Process. The average about k-means was 10.514% and the standard deviation was 0.127%. The MADIC result was 6 SD from the k-means' average. This means that MADIC generated more homogenous clusters than k-means.

Table 2: Distribution of highest keyword.

k-means $N = 80$	10.369%
k-means $N = 90$	10.575%
k-means $N = 100$	10.599%
MADIC	11.445%

5 Conclusion

Traditional clustering algorithms cannot use all the data from repeated measurements. If deviations in repeated measurements are ignored, genes that have big errors can affect the results. Our experiment with 2-dimensional synthetic data shows better results than the k-means algorithm. Our experiment with 16-dimensional synthetic data shows the robustness to errors. In the experiment with real data, we assume MADIC creates the appropriate clusters.

The following features make MADIC a useful method for clustering results from gene expression experiments:

(1) MADIC can handle repeated measurements with error margins; it can identify more stable clusters for stable genes.

(2) The input parameters do not affect the clusters.

(3) Random seed numbers are not needed.

(4) Even if the number of members is one, a peculiar pattern can be extracted as a cluster.

By using our new algorithm, we were able to perform unsupervised clustering of all gene microarray data generated by the Percellome System. This algorithm provides a new option for the analysis of gene expression data.

Our algorithm adopts the gamma function as a density function. Even if an unstable object exists close to a stable object, as a characteristic of the gamma function the unstable object does not affect the stable object. However, because the gamma function does not permit integration by the odd number dimension, it cannot be applied to odd number dimension data. Moreover, it is believed that the device is necessary for very high dimensional data because the smoothness of the incomplete gamma function is lost.

Although our algorithm was designed to analyze microarray data, it should be useful for other types of data that retain error or variation information, and we will subsequently try to apply MADIC to other fields.

Acknowledgments

This work was supported by the Health Sciences Research Grant from the Ministry of Health, Labor and Welfare, Japan, H13-Seikatsu-012, H13-Seikatsu-013, H14-Toxico-001, and H15-Kagaku-002, and in collaboration with the NTT COMWARE CORPORATION.

References

[1] Ester, M., Kriegel, H. P., Sander, J., and Xu, X., A density-based algorithm for discovering clusters in large spatial databases with noise, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231, 1996.

[2] Kanno, J., Aisaki, K., Igarashi, K., *et al.*, "Per cell" normalization method for mRNA measurement by quantitative PCR and microarrays, (in preparation).

[3] Kumar, M., Patel, N. R., and Woo, J., Clustering seasonality patterns in the presence of errors, *Proceedings of the eighth ACM SIGKDD international conference Knowledge Discovery and data mining*, 557–563, 2002.

[4] Monti, S., Tamayo, P., Mesirov, J., and Golub, T., Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data, *Machine Learning*, 52:91–118, 2003.

[5] Papadopoulos, D., Domeniconi, C., Gunopulos, D., and Ma S., Clustering gene expression data in SQL using locally adaptive metrics, *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery (DMKD 2003)*, 235–41, 2003.

[6] Yeung, K. Y., Medvedovic, M., and Bumbgarner, R. E., Clustering gene-expression data with repeated measurements, *Genome Biol.*, 4(5):R34, 2003.

Differential contributions of Mesp1 and Mesp2 to the epithelialization and rostro-caudal patterning of somites

Yu Takahashi[1,*], Satoshi Kitajima[1], Tohru Inoue[1], Jun Kanno[1] and Yumiko Saga[2,*]

[1]Cellular & Molecular Toxicology Division, National Institute of Health Sciences, 1-18-1 Kamiyoga, Setagayaku, Tokyo 158-8501, Japan

[2]Division of Mammalian Development, National Institute of Genetics, Yata 1111, Mishima 411-8540, Japan

*Authors for correspondence (e-mail: yutak@nihs.go.jp and ysaga@lab.nig.ac.jp)

Accepted 29 November 2004

Development 132, 787-796
Published by The Company of Biologists 2005
doi:10.1242/dev.01597

Summary

Mespl and Mesp2 are homologous basic helix-loop-helix (bHLH) transcription factors that are co-expressed in the anterior presomitic mesoderm (PSM) just prior to somite formation. Analysis of possible functional redundancy of Mespl and Mesp2 has been prevented by the early developmental arrest of Mespl/Mesp2 double–null embryos. Here we performed chimera analysis, using either Mesp2-null cells or Mespl/Mesp2 double–null cells, to clarify (1) possible functional redundancy and the relative contributions of both Mespl and Mesp2 to somitogenesis and (2) the level of cell autonomy of Mesp functions for several aspects of somitogenesis. Both Mesp2-null and Mespl/Mesp2 double–null cells failed to form initial segment borders or to acquire rostral properties, confirming that the contribution of Mespl is minor during these events. By contrast, Mespl/Mesp2 double–null cells contributed to neither epithelial somite nor dermomyotome formation, whereas Mesp2-null cells partially contributed to incomplete somites and the dermomyotome. This indicates that Mespl has a significant role in the epithelialization of somitic mesoderm. We found that the roles of the Mesp genes in epithelialization and in the establishment of rostral properties are cell autonomous. However, we also show that epithelial somite formation, with normal rostro-caudal patterning, by wild-type cells was severely disrupted by the presence of Mesp mutant cells, demonstrating non-cell autonomous effects and supporting our previous hypothesis that Mesp2 is responsible for the rostro-caudal patterning process itself in the anterior PSM, via cellular interaction.

Key words: Somitogenesis, Epithelial-mesenchymal conversion, Mesp2, Chimera analysis, Mouse

Introduction

Somitogenesis is not only an attractive example of metameric pattern formation but is also a good model system for the study of morphogenesis, particularly epithelial-mesenchymal interconversion in vertebrate embryos (Gossler and Hrabe de Angelis, 1997; Pourquié, 2001). The primitive streak, or tailbud mesenchyme, supplies the unsegmented paraxial mesoderm, known as presomitic mesoderm (PSM). Mesenchymal cells in the PSM undergo mesenchymal-epithelial conversion to form epithelial somites in a spatially and temporally coordinated manner. Somites then differentiate, in accordance with environmental cues from the surrounding tissues, into dorsal epithelial dermomyotome and ventral mesenchymal sclerotome (Borycki and Emerson, 2000; Fan and Tessier Lavigne, 1994). Hence, the series of events that occur during somitogenesis provide a valuable example of epithelial-mesenchymal conversion. The dermomyotome gives rise to both dermis and skeletal muscle, whereas the sclerotome forms cartilage and bone in both the vertebrae and the ribs. Each somite is subdivided into two compartments, the rostral (anterior) and caudal (posterior) halves. This rostro-caudal polarity appears to be established just prior to somite formation (Saga and Takeda, 2001).

Mespl and Mesp2 are closely related members of the basic helix-loop-helix (bHLH) family of transcription factors but share significant sequence homology only in their bHLH regions (Saga et al., 1996; Saga et al., 1997). During development of the mouse embryo, both *Mespl* and *Mesp2* are specifically expressed in the early mesoderm just after gastrulation and in the paraxial mesoderm during somitogenesis. Mespl/Mesp2 double-null embryos show defects in early mesodermal migration and thus fail to form most of the embryonic mesoderm, leading to developmental arrest (Kitajima et al., 2000). Mespl-null embryos exhibit defects in single heart tube formation, due to a delay in mesodermal migration, but survive to the somitogenesis stage (Saga et al., 1999), suggesting that there is some functional redundancy, i.e. compensatory functions of Mesp2 in early mesoderm. During somitogenesis, both *Mespl* and *Mesp2* are expressed in the anterior PSM just prior to somite formation. Although we have shown that Mesp2, but not Mespl, is essential for somite formation and the rostro-caudal patterning of somites (Saga et al., 1997), a possible functional redundancy between Mespl and Mesp2 has not yet been clearly established.

To further clarify the contributions of Mespl and Mesp2 to somitogenesis, analysis of Mespl/Mesp2 double-null embryos is necessary, but because of the early mesodermal defects already described, these knockout embryos lack a paraxial mesoderm, which prevents any analysis of somitogenesis. We therefore adopted a strategy that utilized chimera analysis. As we have reported previously, the early embryonic lethality of a Mesp1/Mesp2 double knockout is rescued by the presence of wild-type cells in a chimeric embryo, but the double-null cells cannot contribute to the cardiac mesoderm (Kitajima et al., 2000). This analysis, however, focused only on early heart morphogenesis and did not investigate the behavior of Mesp1/Mesp2 double-null cells in somitogenesis. In this report, we focus upon somitogenesis and compare two types of chimeras using either Mesp1/Mesp2 double-null cells or Mesp2-null cells to investigate Mesp1 function during somitogenesis.

Another purpose of our chimera experiments was to elucidate the cell autonomy of Mesp functions. In the process of somite formation, mesenchymal cells in the PSM initially undergo epithelialization at the future segment boundary, independently of the already epithelialized dorsal or ventral margin of the PSM (Sato et al., 2002). Epithelial somite formation is disrupted in the Mesp2-null embryo, indicating that Mesp2 is required for epithelialization at the segment boundary. Although Mesp products are nuclear transcription factors and their primary functions must therefore be cell autonomous (transcriptional control of target genes), it is possible that the roles of Mesp2 in epithelialization are mediated by the non-cell autonomous effects of target genes. We therefore asked whether the defects in Mesp2-null cells during epithelialization could be rescued by the presence of surrounding wild-type cells. Additionally, we would expect to find that the role of Mesp2 in establishing rostro-caudal polarity is rescued in a similar way.

Our analysis suggests that Mesp1 and Mesp2 have redundant functions and are both cell-autonomously involved in the epithelialization of somitic mesoderm. In addition, our results highlight some non-cell autonomous effect of Mesp2-null and Mesp1/Mesp2-null cells.

Materials and methods

Generation of chimeric embryos

As described previously (Kitajima et al.. 2000), chimeric embryos were generated by aggregating 8-cell embryos of wild-type mice (ICR) with those of mutant mice that were genetically marked with the *ROSA26* transgene (Zambrowicz et al.. 1997). Mesp1/Mesp2 double-null embryos were generated by crossing *wko-del (+/–)* and *Mesp1(+/–)/Mesp2(+/cre)* mice as described previously (Kitajima et al.. 2000). This strategy enables us to distinguish chimeric embryos derived from homozygous embryos. which have two different mutant alleles, from those derived from heterozygous embryos. Likewise. Mesp2-null embryos were generated by crossing *P2vl(+/–)* mice (Saga et al.. 1997) and *P2GFP (+/gfp)* mice (Y.S. and S.K.. unpublished) that were also labeled with the *ROSA26* locus. The genotype of the chimeric embryos was determined by PCR using yolk sac DNA.

Histology, histochemistry and gene expression analysis

The chimeric embryos were fixed at 11 days postcoitum (dpc) and stained in X-gal solution for the detection of β-galactosidase activity. as described previously (Saga et al.. 1999). For histology. samples stained by X-gal were postfixed with 4% paraformaldehyde. dehydrated in an ethanol series. embedded in plastic resin (Technovit 8100. Heraeus Kulzer) and sectioned at 3 µm. The methods used for gene expression analysis by in-situ hybridization of whole-mount samples and frozen sections and skeletal preparation by Alcian Blue/Alizarin Red staining were described previously (Saga et al., 1997; Takahashi et al., 2000). Probes for in-situ hybridization for *Uncx4.1* (Mansouri et al.. 1997; Neidhardt et al., 1997). *Delta-like 1* (*Dll1*) (Bettenhausen et al.. 1995) and *Paraxis* (Burgess et al., 1995) were kindly provided by Drs Peter Gruss. Achim Gossler and Alan Rawls, respectively. A probe for *EphA4* (Nieto et al., 1992) was cloned by PCR. For detection of actin filaments. frozen sections were stained with AlexaFluor 488-conjugated phalloidin (Molecular Probes) according to the manufacturer's protocol.

Results

Possible functional redundancy and different contributions of Mesp1 and Mesp2 in somitogenesis

During somitogenesis, both *Mesp1* and *Mesp2* are expressed in the anterior PSM just prior to somite formation and their expression domains overlap (Fig. 1A). Mesp1-null embryos form morphologically normal somites and show normal rostro-caudal patterning within each somite (Fig. 1B,E-H), indicating that Mesp1 is not essential for somitogenesis. By contrast, Mesp2 is essential for both the formation and rostro-caudal patterning of somites, as Mesp2-null embryos have no epithelial somites and lose rostral half properties, resulting in caudalization of the entire somitic mesoderm (Saga et al., 1997) (Fig. 1C,D).

Although somite formation and rostro-caudal patterning is disrupted in the Mesp2-null embryo, histological differentiation into dermomyotome and sclerotome is not affected. It is noteworthy that the Mesp2-null embryo still forms disorganized dermomyotomes without forming epithelial somites (Saga et al., 1997). As *Mesp1* is expressed at normal levels in the PSM of Mesp2-null embryos (Fig. 1C,D), it is possible that Mesp1 functions to rescue some aspects of somitogenesis in the Mesp2-null embryo. In order to further clarify the contributions of both Mesp1 and Mesp2 during somitogenesis, we therefore generated chimeric embryos with either Mesp2-null cells or Mesp1/Mesp2 double-null cells and compared the behavior of mutant cells during somitogenesis (Fig. 2).

Mesp2-null cells tend to be eliminated from the epithelial somite and the dermomyotome, but can partially contribute to both of these structures

We first generated Mesp2-null chimeric embryos (*Mesp2*$^{-/-}$ with *Rosa26*: wild) to analyze cell autonomy of Mesp2 function during somitogenesis. The control chimeric embryo (*Mesp2*$^{+/-}$ with *Rosa26*: wild) showed normal somitogenesis and a random distribution of X-gal stained cells (Fig. 3A). The Mesp2-null chimeric embryos formed abnormal somites that exhibited incomplete segmentation (Fig. 3B), but histological differentiation of dermomyotome and sclerotome was observed. Within the incomplete somite, X-gal-stained Mesp2-null cells were mainly localized in the rostral and central regions, surrounded by wild-type cells at the dorsal, ventral and caudal sides (Fig. 3B). The surrounding wild-type cells, however, did not form an integrated epithelial sheet, but consisted of several epithelial cell clusters. Such trends were more obviously observed in other sections, where wild-type cells were found to form multiple small epithelial clusters (Fig.

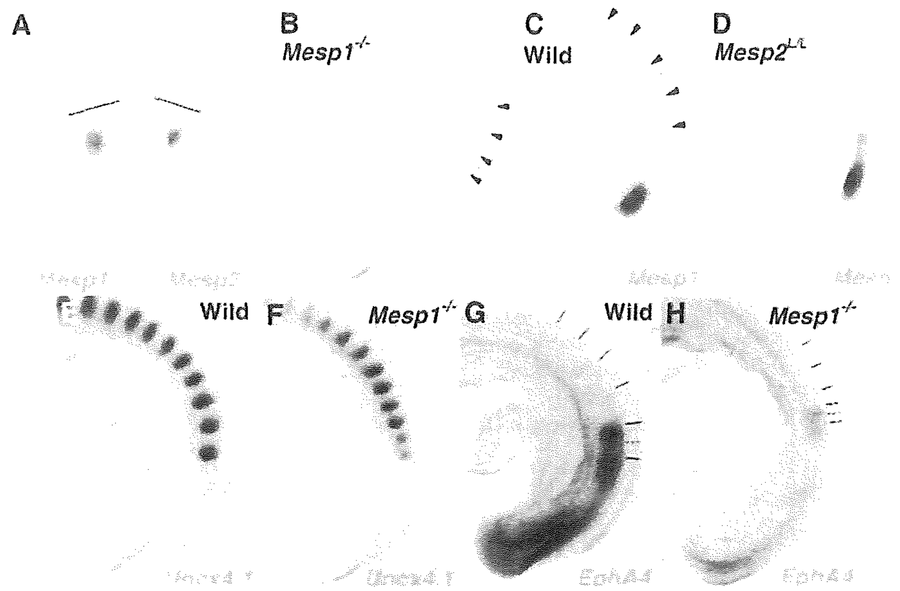

Fig. 1. *Mesp1* and *Mesp2* are co-expressed in the anterior PSM but have differing roles in somitogenesis. (A) Overlapping expression of *Mesp1* and *Mesp2* is revealed by in-situ hybridization using the left and right halves of the same embryo. The lines show most recently formed somite boundaries. (B-C) A Mesp1-null embryo (B) shows the same normal somite formation as a wild-type embryo (C). Arrowheads indicate somite boundaries. (D) In Mesp2-null embryos, no somite formation is observed but *Mesp1* is expressed at comparable levels to wild type, although its expression is anteriorly extended and blurred. (E-H) Mesp1-null embryos show normal rostro-caudal patterning of somites. (E,F) Expression of a caudal half marker, *Uncx4.1*. (G,H) Expression of a rostral half marker, *EphA4*. The lines indicate presumptive or formed somite boundaries and the dotted line indicates approximate position of somite half boundary.

3C,D). Mesp2-null cells tended to be eliminated from the epithelial clusters, although they were partially integrated into these structures (blue arrows in Fig. 3C,D). Likewise, small numbers of Mesp2-null cells were found to contribute to the dermomyotome (Fig. 3E,F). Mesp2-null cells also appeared to form the major part of the sclerotome.

Mesp2 is required for the cell-autonomous acquisition of rostral properties

We have previously demonstrated that suppression by Mesp2 of the caudal genes *Dll1* and *Uncx4.1* in presumptive rostral half somites is a crucial event in the establishment of the rostro-caudal pattern of somites (Saga et al., 1997; Takahashi et al., 2000). As Mesp2-null embryos exhibit caudalization of somites, Mesp2-null cells are predicted to be unable to express rostral properties. Hence, Mesp2-null cells are expected to distribute to the caudal region of each somite where the rostro-caudal patterns are rescued by wild-type cells in a chimeric embryo. In this context, the localization of Mesp2-null cells at the rostral side was an unexpected finding. We interpret this to mean that the rostral location of Mesp2-null cells is due to a lack of epithelialization functions (see Discussion).

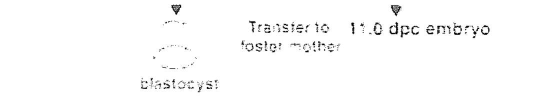

Fig. 2. Schematic representation of chimera analysis method. Either Mesp2-null or Mesp1/Mesp2 double-null embryos, genetically labeled with *Rosa* locus, were aggregated with wild-type embryos at the 8-cell stage, and the resulting chimeras were subjected to analysis at 11.0 dpc.

To examine rostro-caudal properties in Mesp2-null cells, located in the rostral side, we analyzed the expression of a caudal half marker gene, *Uncx4.1* (Mansouri et al., 1997; Neidhardt et al., 1997). Analysis of adjacent sections revealed that *lacZ*-expressing Mesp2-null cells, localized at the rostral and central portion, ectopically expressed *Uncx4.1* (Fig. 4A-D). This strongly suggests that Mesp2-null cells cannot acquire rostral properties even if surrounded by wild-type cells, and that Mesp2 function is cell-autonomously required for the acquisition of rostral properties. We also observed that the small number of Mesp2-null cells distributed mostly to the caudal end of the dermomyotome (Fig. 3E,F) and that the expression pattern of *Uncx4.1* was normal in the dermomyotome (Fig. 4E,F). In the sclerotome, *lacZ*-expressing Mesp2-null cells often distributed to the rostral side, where expression of *Uncx4.1* was abnormally elevated (Fig. 4G,H). The vertebrae of the Mesp2-null chimeric fetus showed a partial fusion of the neural arches, which was reminiscent of