$$D_2(g,S_i) = -\log_{10}p \quad (4)$$

Regarding each SQ specimen S_i (i = 1, 2,. . ., 21), the probes whose differential levels $D_2(g,S_i)$ were equal to or more than *diff* were defined as the individual-specimen cluster, $C_{sign\_diff\_Si}$, where *sign* is the differential direction (+, overexpression; −, underexpression in each SQ specimen). $C_{sign\_diff\_Si}$ was defined for all

$$sign = -, +$$
$$diff = 2, 3, 4, \ldots$$
$$S_i = 1, 2, \ldots, 21$$

For example, C_{+2_Si} and C_{-2_Si} were clusters of probes whose expression of S_i were included in 1% of sections on both sides of NL's distributions. More specifically, C_{+2_Si} was a cluster of probes whose expression levels were equal to or higher than (ave_{NL} + 2.58 $stddev_{NL}$) in a specimen S_i, where ave_{NL} is the mean and $stddev_{NL}$ is the standard deviation of expression level in NL specimens. In the same manner, C_{-2_Si} was a cluster of probes whose expression levels were equal to or less than (ave_{NL} − 2.58 $stddev_{NL}$); $n_{sign\_diff\_Si}$ is the number of Key-UniGenes in $C_{sign\_diff\_Si}$. If multiple probes in a cluster could be mapped to single UniGene, then only the probe with the highest D_2 value was adopted. The average numbers, $\bar{n}_{sign\_diff}$, of ($n_{sign\_diff\_Si}$)(i = 1, 2,. . ., 21) are shown in Table 3.

Construction of the EIM. In a manner similar to the EIM for detecting expression imbalance of SQ group, that for detecting individual differences in expression imbalance among SQs was also constructed. The individual-specimen clusters, $C_{sign\_diff\_Si}$, were arranged on the abscissa with respect to each S_i, and the locus clusters on the ordinate (Fig. 8). Underexpression clusters were arranged on the left side and overexpression clusters on the right. Since the abscissa represented an array of S_i, it was impossible to represent *diff* on the abscissa like Fig. 4. Therefore, the EIM for individual specimen was visualized by $C_{sign\_diff\_Si}$ with a defined *diff*, and allowed the user to change *diff* interactively.

The number of common Key-UniGenes between $C_{sign\_diff\_Si}$ and $C_{arm\_length\_begin}$, k, could also be evaluated using $E(U, n_1, n_2, k)$ (*Eq. 3*), where n_1 was $\bar{n}_{sign\_diff}$ and n_2 was $n_{arm\_length\_begin}$. If the different specimens have the same number of genes with under- or overexpression on the same local region, then it is necessary to evaluate them as similar. Therefore, $\bar{n}_{sign\_diff}$ instead of $n_{sign\_diff\_Si}$ was used for the evaluation of the overlap between $C_{sign\_diff\_Si}$ and $C_{arm\_length\_begin}$. The E value for any combination of $C_{sign\_diff\_Si}$ and $C_{arm\_length\_begin}$ was calculated,

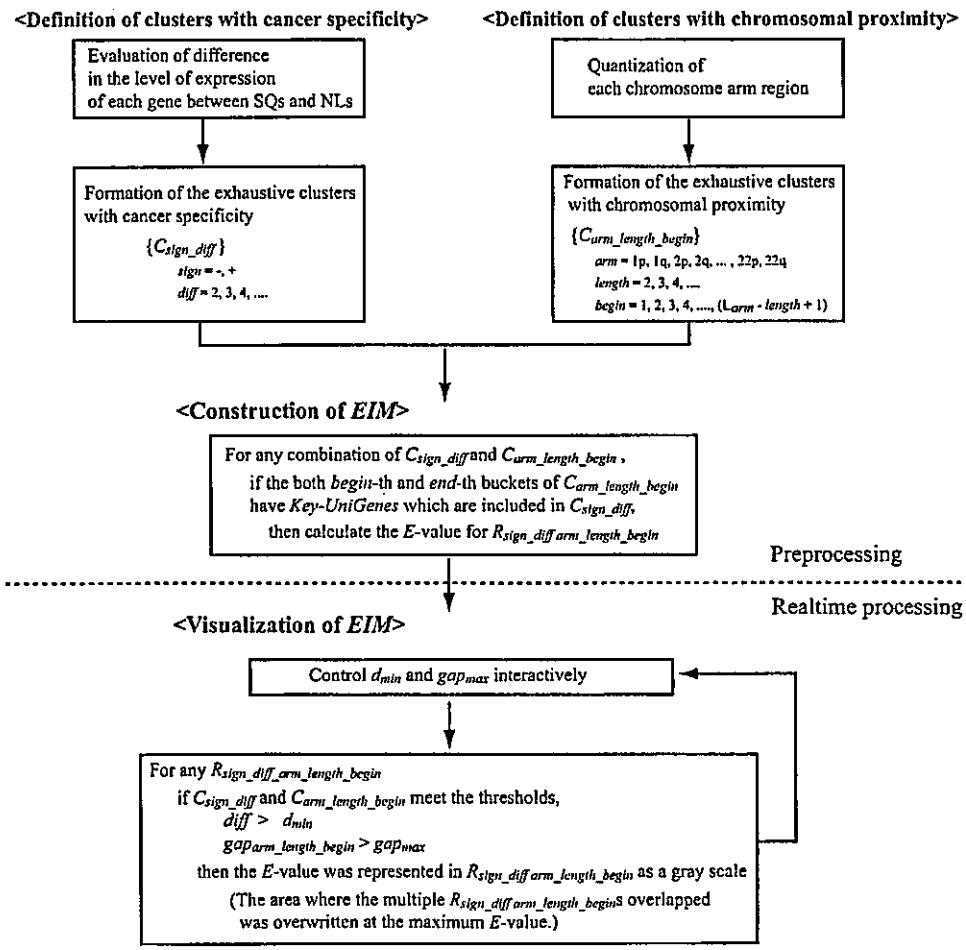


Fig. 5. Flowchart for construction of the EIM for detecting expression imbalance regions specific to SQs. This flowchart provides details of the steps of the EIM for detecting expression imbalance regions specific to SQs. For the steps of "Definition of clusters with cancer specificity," please refer to Fig. 3. For the steps of "Definition of clusters with chromosomal proximity," please refer to Fig. 2. For the steps of "Construction of the EIM" and "Visualization of EIM," please refer to Fig. 4. The user can interactively control the steps in real-time processing by changing gap_{max} and d_{min}.

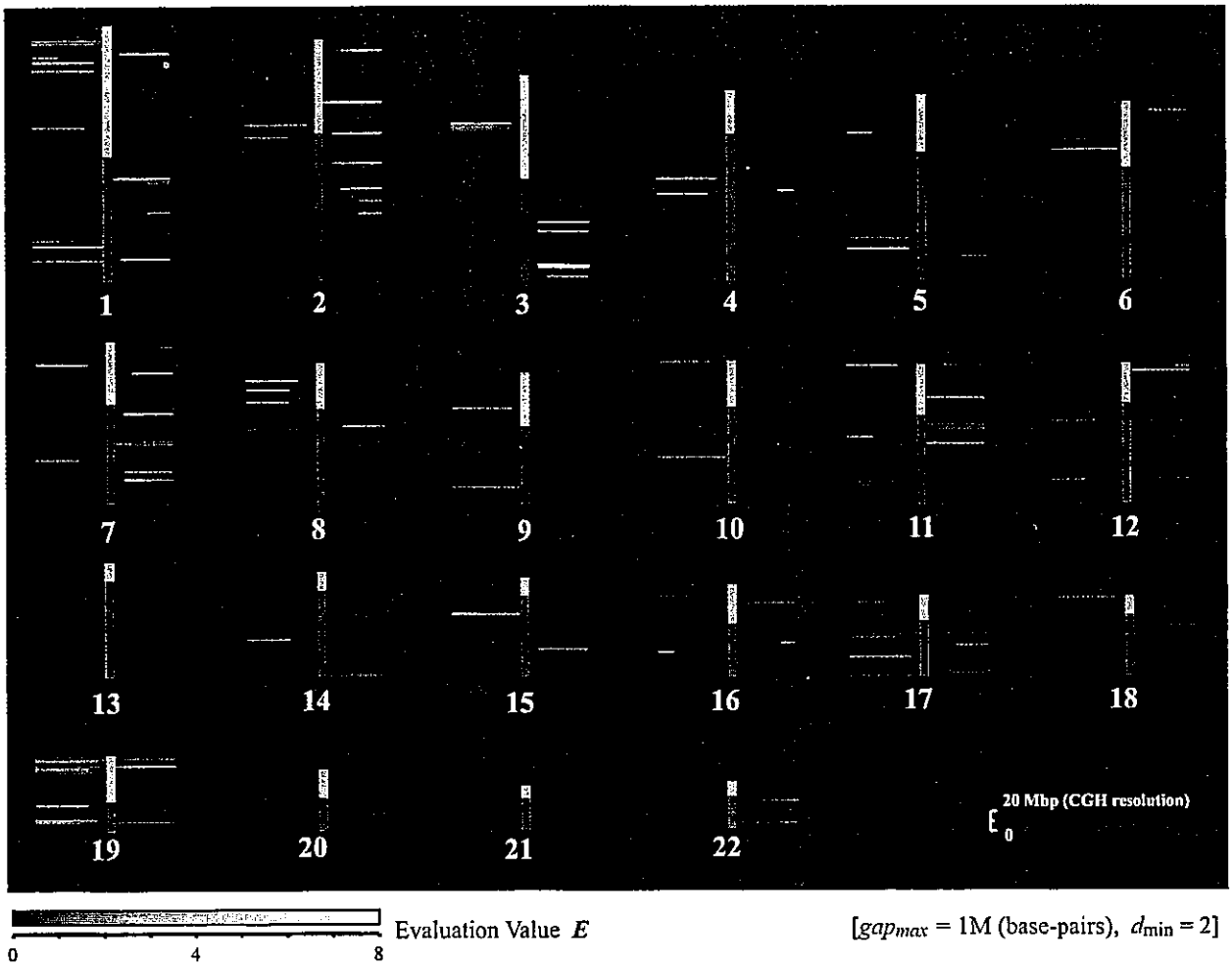


Fig. 6. The EIM applied for detecting expression imbalance regions specific to SQs. The regions of under- and overexpression in SQs were visualized on the *left* and *right* side, respectively, as gray regional signals. All statistical evaluation values of any combinations between the exhaustive uncertainty cluster sets of cancer specificity and chromosomal proximity are visualized on the EIM as the gradation of gray scale simultaneously. Each exhaustive uncertainty cluster set was formed by repetition of the sufficiently minute changes of the threshold of cancer specificity or chromosomal proximity. While the area with high luminance corresponds to the more probable expression imbalance region, the EIM enables the user to search as many genes as possible by referring to more expanded area with lower luminance. The EIM presented the most significant overexpression regions on 3q (the evaluation value $E = 7.2$), which is a well-known locus with frequent genomic gains, as detected by comparative genomic hybridization (CGH) (6, 8, 9). Note the high resolution of the EIM compared with CGH resolution (~20 Mbp).

Fig. 7. Expression imbalance regions specific to SQs on chromosome 3. *A–I*: chromosome 3 of the EIM and the influence of gap_{max} and d_{min} on the detection of the expression imbalance regions specific to SQs. The EIM represents the E values whose C_{sign_diff} and $C_{arm\_length\_begin}$ meet d_{min} and gap_{max}, respectively. The EIM allows the user to control gap_{max} and d_{min} interactively. The user can narrow down the possible expression imbalance regions by changing gap_{max} and d_{min}. Especially, as is shown in *A–I*, changing gap_{max}, which allows exclusion of regions containing large gaps between genes, markedly affected the detection of expression imbalance regions. *J*: the macrograph of the encircled *region A* from panel *A*. Intersection area $R_{+5\_3q\_1894\_5}$ shows the most significant overexpression region, which is a well-known locus with frequent genomic gains as previously detected by CGH (6, 8, 9). That is, the overlap ($k = 6$) between C_{+5} and $C_{3q\_1894\_5}$ was statistically the most significant ($E = 7.2$). C_{+5} was the cluster of probes with overexpression whose differential level $D_1(g)$ was more than 5 and its number of Key-UniGenes, n_{+5}, was 205. $C_{3q\_1894\_5}$ was the region from 189,400 to 189,900 kbp on chromosome 3 and contained 9 Key-UniGenes ($n_{3q\_1894\_5} = 9$). The maximum gap ($gap_{3q\_1894\_5}$) between Key-UniGenes in $C_{3q\_1894\_5}$ was 146 kbp. In addition, all evaluation values of any combinations between the exhaustive uncertainty cluster sets of cancer specificity and chromosomal proximity are visualized simultaneously on the EIM as gradation of the gray scale. This gradation pattern could convey the distribution of the false balance to the user through visual perception and enabled the detection of as many significant genes as possible. In addition, note the high resolution of EIM compared with CGH resolution (~20 Mbp).

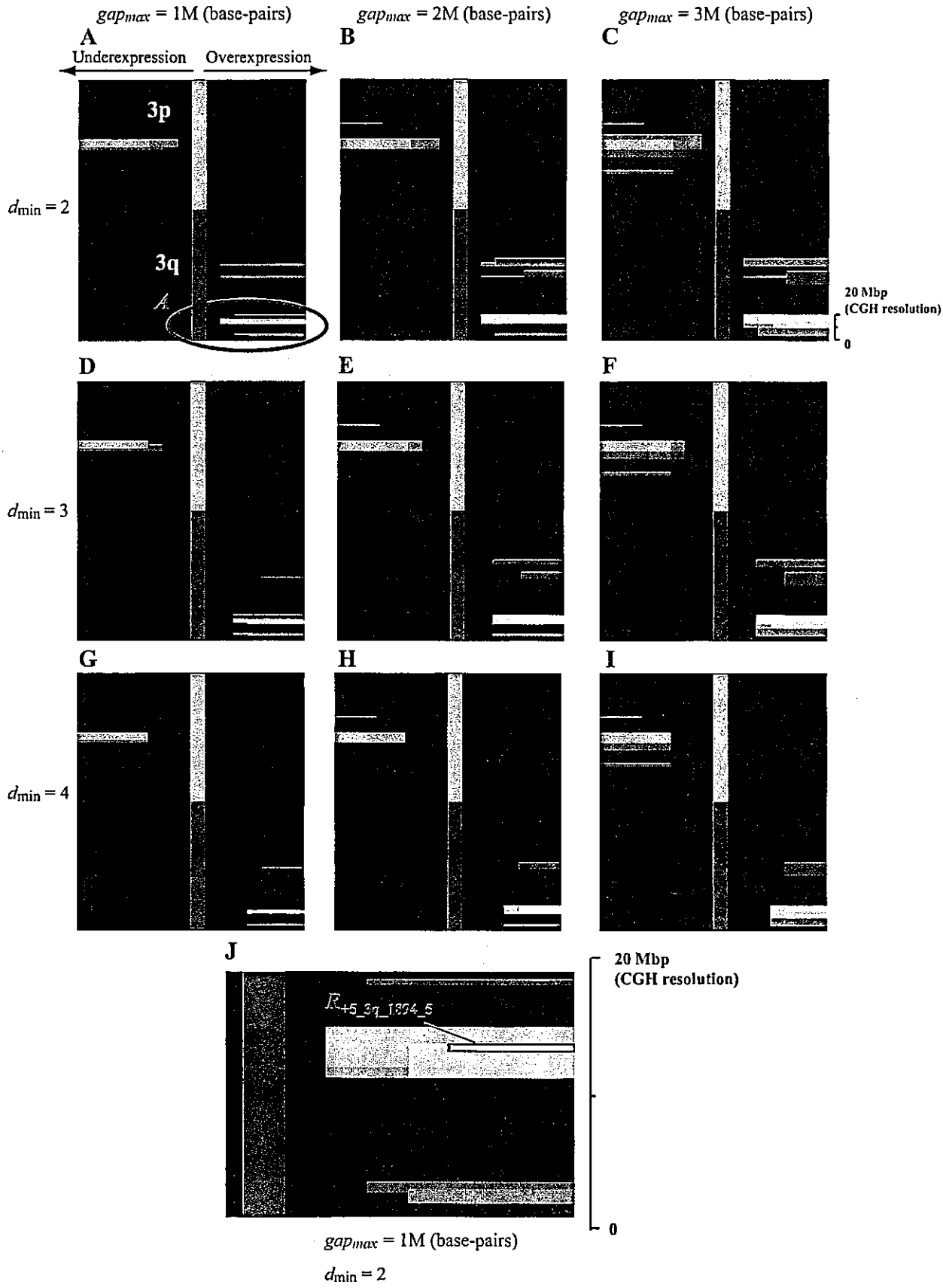

gap_{max} = 1M (base-pairs)
gap_{max} = 2M (base-pairs)
gap_{max} = 3M (base-pairs)
A
Underexpression
Overexpression
B
C
3p
3q
d_{min} = 2
20 Mbp
(CGH resolution)
0
D
E
F
d_{min} = 3
G
H
I
d_{min} = 4
J
R+5_3q_1894_5
20 Mbp
(CGH resolution)
0
gap_{max} = 1M (base-pairs)
d_{min} = 2

Table 3. *Clusters of probes with under- or overexpression profiles in each squamous cell lung carcinoma*

Differential Direction	Cluster Name ($C_{sign\_diff\_Si}$)	Avg. of Probe Number	Avg. of Key-UniGene Number ($\bar{n}_{sign\_diff}$)	SD of Key-UniGene Number
NL(17) > each SQ	C_{-2_Si}	669	447	103
	C_{-3_Si}	497	331	91
	C_{-4_Si}	387	259	82
	C_{-5_Si}	317	211	76
	C_{-6_Si}	268	181	70
NL(17) < each SQ	C_{+2_Si}	321	208	67
	C_{+3_Si}	188	120	48
	C_{+4_Si}	120	77	35
	C_{+5_Si}	81	50	25
	C_{+6_Si}	58	36	19

To detect individual differences in expression imbalance among 21 SQs, probes (on the U95A array) with under- or overexpression profiles in a SQ specimen, S_i (i = 1,2,. . .,21), compared with NLs were extracted as clusters, $C_{sign\_diff\_Si}$. This extraction was independently performed, regarding each SQ specimen. The suffix *sign* indicates the differential direction (+, overexpression; −, underexpression in each SQ specimen), *diff* indicates a differential level D_2 in gene expression. Shown are the average number of probes and the average and standard deviation (SD) of Key-UniGenes in the 21 clusters with the same differential direction and differential level.

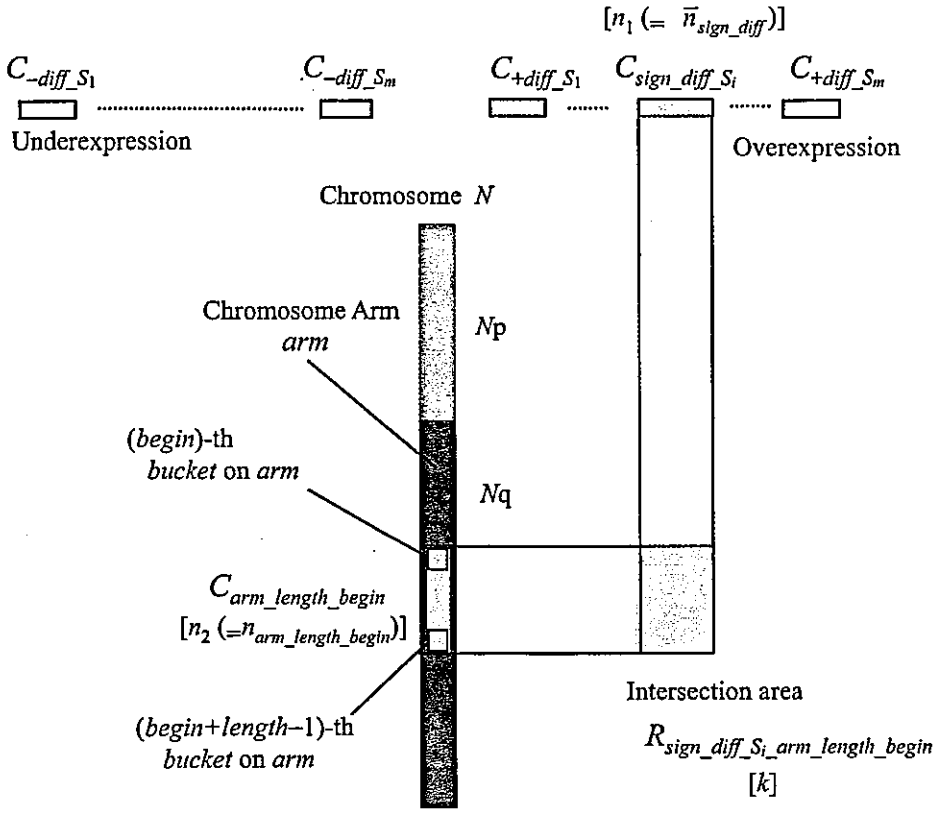


Fig. 8. Individual-specimen clusters vs. locus clusters. In a manner similar to the EIM for detecting expression imbalance of SQ specimen group, that for detecting individual differences in expression imbalance among SQ specimens was also constructed. In a SQ specimen S_i (i = 1, 2,. . ., 21), probes with expression whose differential level $D_2(g,S_i)$ was equal to or higher than *diff* compared with NL specimens were extracted as an individual-specimen cluster, $C_{sign\_diff\_Si}$. This extraction was independently performed with respect to each SQ specimen. The individual-specimen clusters, $C_{sign\_diff\_Si}$ values, were arranged on the abscissa with respect to each S_i, and the locus clusters, $C_{arm\_length\_begin}$ values, on the ordinate. Among $C_{sign\_diff\_Si}$ values, the clusters of under- and overexpression were arranged on the *left* and *right* side, respectively. Since the abscissa represented an array of S_i, it was impossible to represent *diff* on the abscissa like Fig. 4. Therefore, the EIM for individual specimen was visualized by $C_{sign\_diff\_Si}$ with a defined *diff*, and allowed the user to change *diff* interactively; $\bar{n}_{sign\_diff}$ is the average number of Key-UniGenes in $\{C_{sign\_diff\_Si}\}(i = 1, 2,. . ., 21)$; $n_{arm\_length\_begin}$ is the number of Key-UniGenes in $C_{arm\_length\_begin}$; k is the number of common Key-UniGenes between $C_{sign\_diff\_Si}$ and $C_{arm\_length\_begin}$. The significance of overlap between $C_{sign\_diff\_Si}$ and $C_{arm\_length\_begin}$ was visualized in the intersection area $R_{sign\_diff\_Si\_arm\_length\_begin}$ as a gray scale.

when both (*begin*)-th and (*begin* + *length* − 1)-th buckets of $C_{arm\_length\_begin}$ have the Key-UniGenes that are included in $C_{sign\_diff\_Si}$. This calculation was preprocessing for the EIM. Then, in real-time processing, after a certain *diff* was selected, each *E* value was represented in the intersection area, $R_{sign\_diff\_Si\_arm\_length\_begin}$, as a gray scale, if $C_{arm\_length\_begin}$ met gap_{max}. The user can control *diff* and gap_{max} interactively.

A flowchart that details these steps is shown in Fig. 9. The EIM for detecting individual difference of expression imbalance among SQ specimens is shown in Fig. 10. Figure 11 shows chromosome 3 of the EIM and the influence of gap_{max} and *diff* on the detection of the individual differences in expression imbalance among SQs.

RESULTS AND DISCUSSION

Detection of Expression Imbalance Specific to SQs

The EIM showed the distribution of expression imbalance specific to SQs (Fig. 6). It is highly comparable to previous CGH data of lung cancer reported by other investigators (6, 8, 9). There are significant differences among these CGH data because of method variation and sample preparation (especially tumor fraction of clinical samples). So it may be of little importance to compare details with individual CGH experiments. However, the most frequent abnormal loci reported in most of these studies were also detected by the EIM as regional signal images on chromosomes (expression imbalance regions), such as loss of 3p, 4q, 5q, and 8p, and gain of 1q, 3q, and 12p (6, 8, 9). The major difference from the CGH image is that signals are detected in a more confined area, which reflects the high resolution of EIM. Figures 6, 7, 10, and 11 clearly show the high resolution of EIM compared with CGH image. Especially, the intersection area $R_{+5\_3q\_1894\_5}$ showed the most significant overexpression region on 3q (Fig. 7), which is reported to be the most frequent aberration

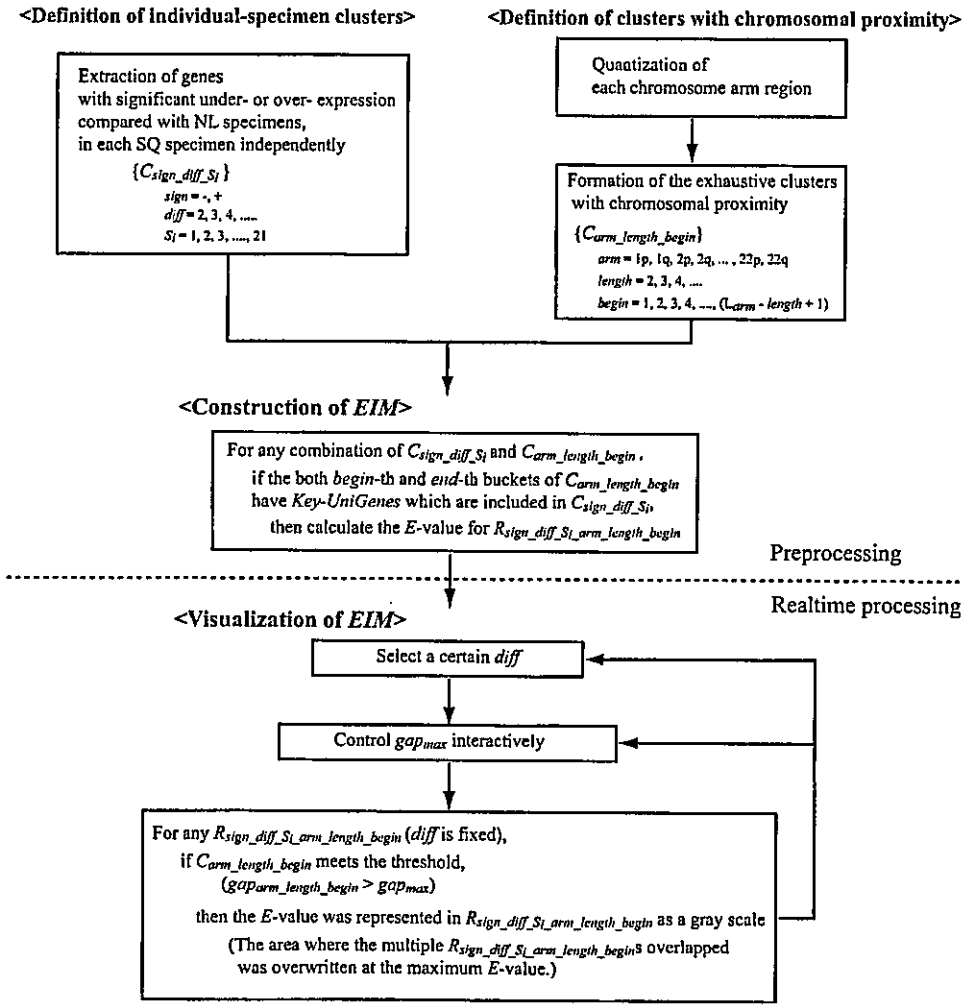


Fig. 9. Flowchart for construction of the EIM for detecting individual differences in expression imbalance among SQs. This flowchart provides details of the steps of the EIM for detecting individual differences in expression imbalance among SQs. For the step of "Definition of clusters with chromosomal proximity," please refer to Fig. 2. For the step of "Construction of the EIM" and "Visualization of EIM," please refer to Fig. 8. In this type of EIM, since the abscissa represented an array of S_i, it was impossible to represent *diff* on the abscissa like Fig. 4. Therefore, the EIM for individual specimen was visualized by $C_{sign\_diff\_Si}$ with a defined *diff*, and allowed the user to change *diff* interactively. In addition, it is possible to exclude regions containing large gaps between genes by changing gap_{max} interactively.

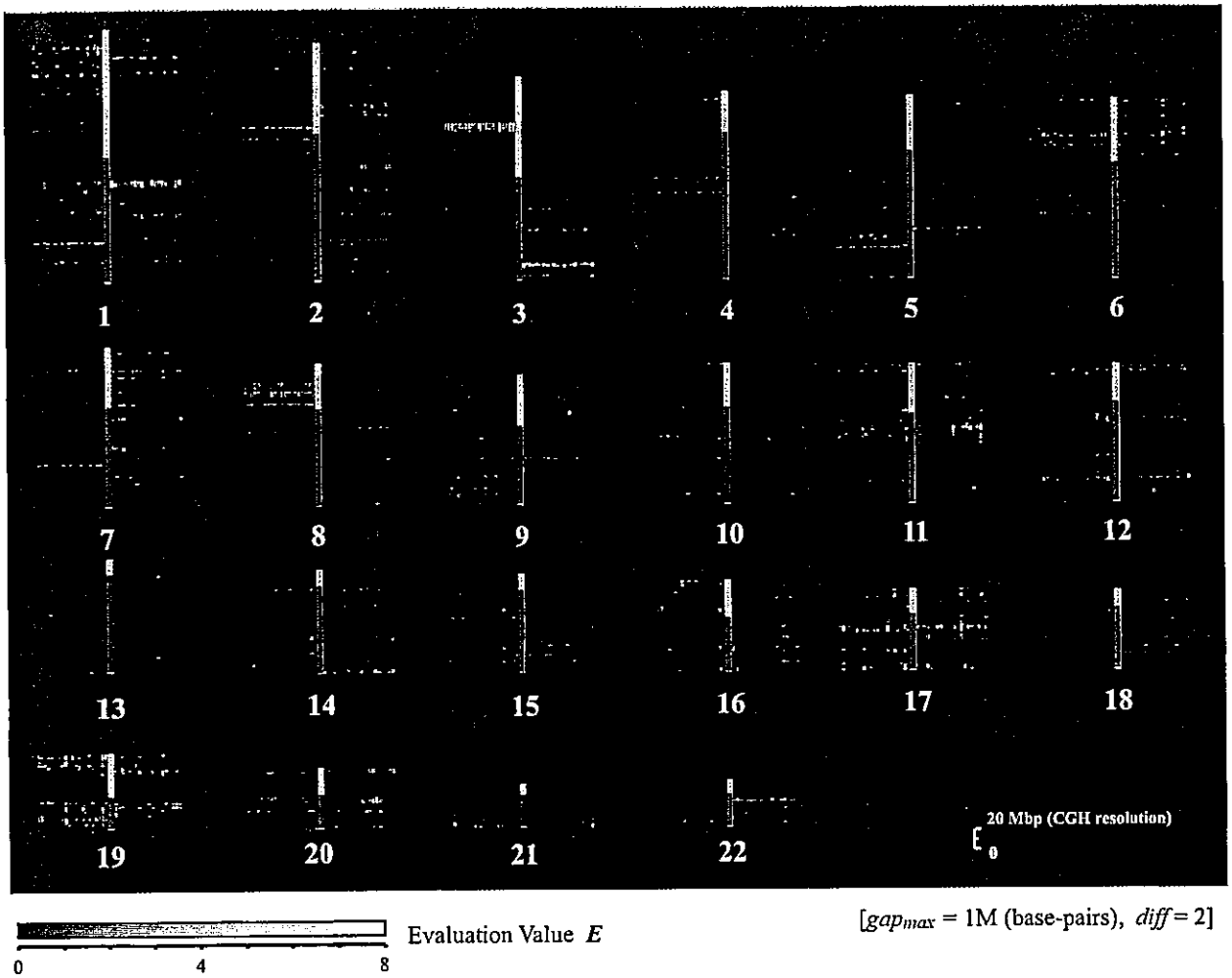
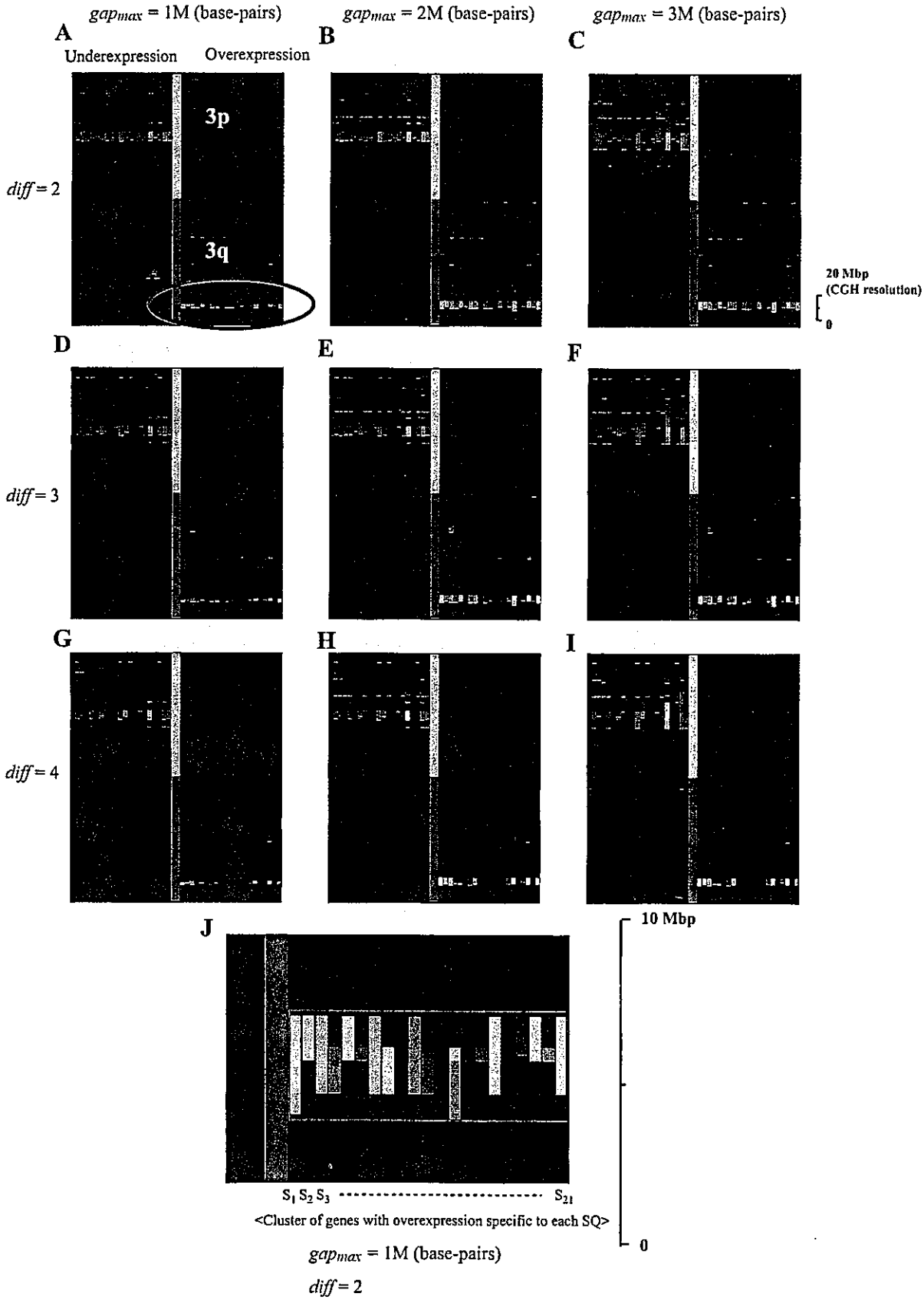


Fig. 10. The EIM for detecting individual difference of expression imbalance among SQs. The EIM was applied for detecting individual differences of expression imbalance among the SQs. Regions of underexpression and overexpression were visualized on the *left* and *right* side, respectively, as gray regional signals. The expression imbalance regions in each SQ were evaluated independently. Note the high resolution of EIM compared with CGH resolution (~20 Mbp).

in SQs by CGH (6, 8, 9). That is, the overlap (k = 6) between C_{+5} (the cluster of probes with overexpression whose differential level $D_1(g)$ is more than 5: n_{+5} = 205) and $C_{3q\_1894\_5}$ (the region from 189,400 to 189,900 kbp on chromosome 3: $n_{3q\_1894\_5} = 9$, $gap_{3q\_1894\_5} = 146$ kbp) was statistically the most significant (E = 7.2). Therefore, the overlap was evaluated using the hypergeometric probability for observing at least 6 (=k) common elements between randomly selected 205 (=n_{+5}) and 9 (=$n_{3q\_1894\_5}$) elements among 6,652 (=U) elements. The user can narrow down the possible expression imbalance regions by changing $gap_{\max}$ and $d_{\min}$ interactively. Especially, as is shown in Fig. 7, *A–I*, changing $gap_{\max}$, which allows exclusion of the regions containing large gaps between genes, markedly influenced the detection of expression imbalance regions. In addition, all evaluation values of any combinations between the exhaustive uncertainty cluster sets of cancer specificity and chromosomal proximity are visualized simultaneously on the EIM as gradation

Fig. 11. Individual difference of expression imbalance on chromosome 3. *A–I*: chromosome 3 of the EIM and the influence of $gap_{\max}$ and *diff* on the detection of individual differences in expression imbalance among SQs. With regard to each SQ specimen, the under- and overexpression regions were visualized on the *left* and *right* side, respectively. Since the expression imbalance regions in each SQ were evaluated independently, this type of EIM clarified the individual difference of the overexpression region on 3q, which was detected as the most significant region in the group of SQs by another type of EIM. The user can narrow down the possible expression imbalance regions by changing $gap_{\max}$ and *diff*. *J*: macrograph of the encircled *region A* from panel *A*. When $gap_{\max}$ was 1 Mbp and *diff* was 2, the EIM showed that 17 of 21 SQs had overexpression regions on 3q, which is comparable to other data sets by CGH (6, 8, 9). In addition, note the high resolution of the EIM compared with CGH resolution (~20 Mbp).


gap_{max} = 1M (base-pairs)
gap_{max} = 2M (base-pairs)
gap_{max} = 3M (base-pairs)
A
B
C
Underexpression
Overexpression
3p
3q
diff = 2
20 Mbp
(CGH resolution)
0
D
E
F
diff = 3
G
H
I
diff = 4
10 Mbp
J
S_1 S_2 S_3 ---------- S_{21}
<Cluster of genes with overexpression specific to each SQ>
0
gap_{max} = 1M (base-pairs)
diff = 2

of gray scale, which is clearly shown in Fig. 7*J*. This gradation pattern could convey the distribution of the false balance to the user through visual perception and enabled the detection of as many significant genes as possible.

Table 4 shows the gene list of $C_{3q\_1894\_5}$. Although this overexpression region strongly reflected the known genomic gain detected by CGH, several probes without overexpression were also detected on this region. There may be several reasons for this. First, since several probes with low quality were possibly included in this region, signal intensity does not always reflect their target mRNA expression levels. Improvement of the quality of probes would make it possible to detect the overexpression region more clearly. Second, mRNA expression levels would not completely reflect genomic copy number changes caused by chromosomal gain or loss, although there was strong correlation between them, because they are under various transcriptional control including feedback pathway of lost or gained genes themselves. Mukasa et al. (7) also reported that several genes without reduction of expression were detected in 1pLOH region of oligodendrogliomas. In addition, it should be stated that cancer tissues used here contained significant number of noncancerous stromal or inflammatory cells, which add noisy expression to cancer profiling.

Because of the complex factors discussed above, simple spatial mapping of the microarray expression profiles on chromosomal location gives little information about genomic structure (Fig. 12, *left*). In addition, it is very difficult to define adequate thresholds for cancer specificity and chromosomal proximity, because the distribution of "false balance" is unclear and the risk of overlooking significant genes by arbitrary selection of thresholds is high (i.e., the "threshold problem"). However, the EIM, using a new methodology without arbitrary selection of thresholds in conjunction with hypergeometric distribution-based algorithm, has a high tolerance of these complex factors and controls the risk of overlooking the expression imbalance regions. This advantage of the EIM over the simple spatial mapping is clearly shown in Fig. 12. The EIM detected the underexpression regions, *A* and *B*, and overexpression region, *C*, on chromosome 11, which are known loci with frequent genomic gain or genomic loss (6, 8, 9), although it was difficult to detect it from the simple spatial mapping of D_1 value.

Detection of Individual Difference in Expression Imbalance Among SQ Specimens

The analysis for extraction of probes with expression profiles specific to the group of cancer is very effective and popular. However, this type of analysis sometimes raises a critical problem because the individual difference among a group is unobservable. In this context, the function of the EIM to detect individual difference of expression imbalance in a group is very significant. Figure 11, *A–I*, shows that the user can narrow down the possible expression imbalance regions on chromosome 3 by changing gap_{max} and *diff* interactively. Furthermore, Fig. 11*J* shows the individual difference in the most significant overexpression regions on 3q (gap_{max} = 1 Mbp, *diff* = 2), where 17 of 21 SQs had overexpression regions, a finding comparable with other data sets analyzed by CGH (6, 8, 9).

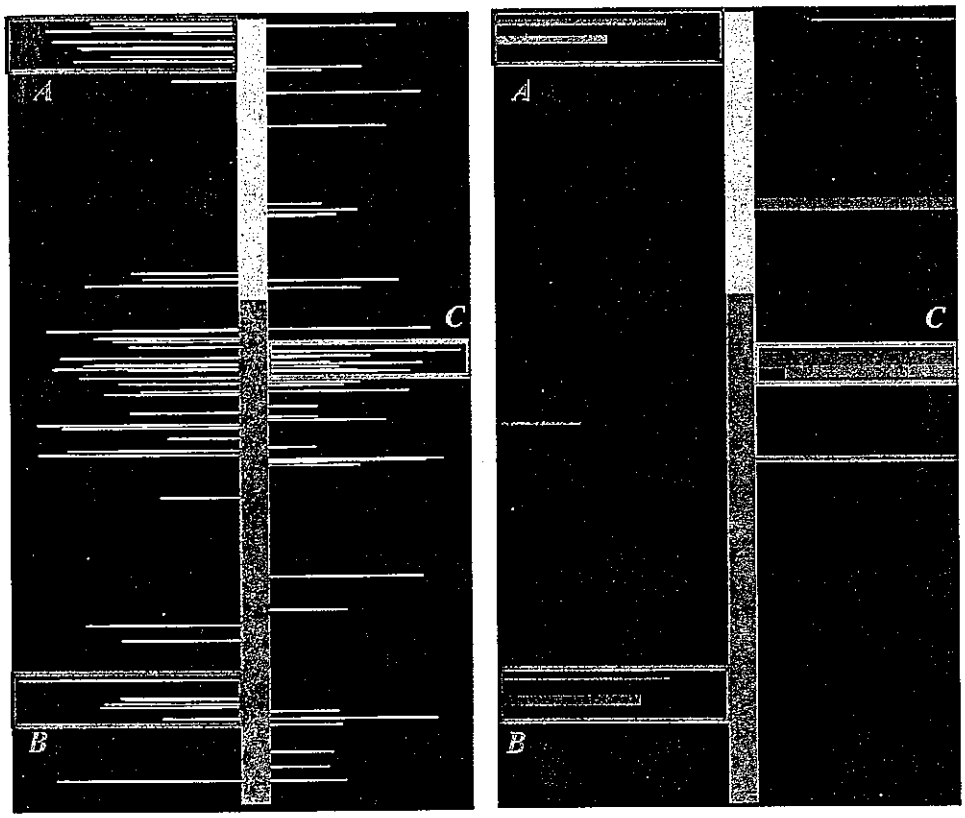
The high-resolution spatial map of expression profiles described in this report, i.e., the EIM, has several significant advantages. Its validity is clearly shown by the fact that many known loci with high frequent genomic losses or gains were detected by regional signals obtained with high resolution by this method.

Recently, several studies have been reported on microarray-based CGH for detecting genome-wide copy number changes (10). However, to our knowledge, no spatial mapping data obtained with such validity and genome-wide coverage have ever been reported previously from this array-CGH method. Experimental difficulty of genome hybridization and limited number of

Table 4. *Gene list of the overexpression region on 3q detected by the EIM*

Cancer Specificity	UniGene	Location, base pairs	Description
*	Hs.108660	189457995	ATP-binding cassette, subfamily C (CFTR/MRP), member_5
?	Hs.343882	189554055	CaM-KII inhibitory protein
x	Hs.129801	189604044	KIAA0604 gene product
x	Hs.1166	189609401	thrombopoietin (myeloproliferative leukemia virus oncogene ligand, megakaryocyte growth and development factor)
*	Hs.74619	189621219	proteasome (prosome, macropain) 26S subunit, non-ATPase, 2
x	Hs.141660	189658124	chloride channel 2
*	Hs.211568	189734699	eukaryotic translation initiation factor 4 gamma, 1
?	Hs.146161	189735389	hypothetical protein MGC2408
*	Hs.153591	189832147	Not56 (*D. melanogaster*)-like protein
*	Hs.174044	189851048	dishevelled 3 (homologous to *Drosophila* dsh)
*	Hs.152936	189862279	adaptor-related protein complex 2, mu 1 subunit

The expression imbalance map (EIM) detected the most significant overexpression regions, $R_{+5\_3q\_1894\_5}$, on 3q in the SQs. This region is a known locus with frequent genomic gains (6, 8, 9). This table shows the gene list of intersection area $R_{+5\_3q\_1894\_5}$. $R_{+5\_3q\_1894\_5}$ evaluated the overlap between C_{+5} (the cluster of probes on the U95A oligonucleotide arrays with overexpression whose differential level are more than 5) and $C_{3q\_1894\_5}$ (the region from 189,400 to 189,900 kbp on chromosome 3: $gap_{3q\_1894\_5}$ = 146 kbp). Differential levels of the genes marked with an asterisk (*) were more than 5, and those of the genes with "x" were less than 5. The genes with "?" were not the Key-UniGenes but the UniGenes that were contained in Genes On Sequence Map.



Simple spatial mapping

Expression Imbalance Map
(gap_{max} = 1M (base-pairs), d_{min} = 2)

Fig. 12. Advantages of the EIM over the simple spatial mapping of expression profiles. *Left*: a simple spatial mapping of D_1 value, which was calculated from the expression profiles of SQs, on chromosome 11. *Right*: the EIM of the same region. The EIM allowed detection of the underexpression regions, *A* and *B*, and overexpression region, *C*, on chromosome 11, which are known loci with genomic gain or genomic loss (6, 8, 9), although it is difficult to detect it by simple spatial mapping.

probes on CGH array could be major problems for it. There may be several reasons for the successful result of our alternative approach, calculation of genomic structure from expression profile. The first reason is the use of the Affymetrix-type GeneChip. The large number of probes (12,533) available enables detection of a relatively short abnormal region (chromosomal loss can frequently affect areas as short as a few hundred kbp), although this method can be easily applied to other types of microarrays. The second reason, which is most important, is that the EIM is a visualization method using a new methodology without arbitrary selection of thresholds in conjunction with hypergeometric distribution-based algorithm. By processing the complex factors and the threshold problems which hinder user's visual perception of essential information, the EIM presents to the user a comprehensive visual image of whole genome-wide information, clearly indicating where expression imbalance regions are and which genes are to be examined. It has an obvious advantage over simple spatial mapping of the expression profiles. For further curation by the user, simple clicking of a selected expression imbalance region on the EIM image leads to a direct link to a file that contains the actual gene names of the region, their expression scores, and other biological information. In addition, if the user input the UniGene number of genes of interest, the EIM indicates its position on the chromosome. Therefore, the EIM can be a broadband interface that enables user's visual perception of complex data and further curation.

Using the EIM, we might be able to detect regional under- or overexpressions independent of copy number changes, such as gene methylation silencing and/or imprinting abnormality (11). In addition, by using the Kruskal-Wallis test (4), which is a rank sum test to deal with three or more data groups instead of Mann-Whitney test, the EIM can easily extend to multiple phenotypes.

In conjunction with the microdissection technique, which can isolate only tumor-cell-specific RNA (2), our EIM can more precisely detect potential genomic structural changes, which offer more diagnostic and therapeutic impact.

Conclusion

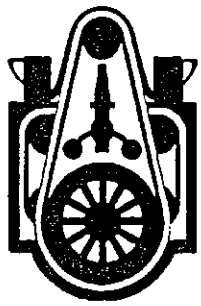
In this report, we describe the development of the expression imbalance map, or EIM, a visualization method without arbitrary selection of thresholds, in conjunction with hypergeometric distribution-based algorithm, for detecting expression imbalance regions. By using this method, many known as well as potential loci with high frequent genomic losses or gains were detected as regional signals with much higher resolution than conventional methods, such as CGH. The EIM can be a broadband interface which enables user's visual perception of complex data and further curation,

and its advantage is obvious over simple spatial mapping of the expression profiles on chromosomal location. Therefore, the EIM would provide the user with further insight into the genomic structure through mRNA expression.

This work was supported by Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES

1. **Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, and Meyerson M.** Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 98: 13790–13795, 2001.
2. **Bonner RF, Emmert-Buck M, Cole K, Pohida T, Chuaqui R, Goldstein S, and Liotta LA.** Laser capture microdissection: molecular analysis of tissue. *Science* 278: 1481–1483, 1997.
3. **Fujii T, Dracheva T, Player A, Chacko S, Clifford R, Strausberg LS, Buetow K, Azumi N, Travis WD, and Jen J.** A preliminary transcriptome map of non-small cell lung cancer. *Cancer Res* 62: 3340–3346, 2002.
4. **Hayter AJ.** *Probability and Statistics for Engineers and Scientists* (2nd ed.). Florence, KY: Duxbury Press, 2002.
5. **Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, and Pinkel D.** Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 258: 818–821, 1992.
6. **Lu YJ, Dong XY, Shipley J, Zhang RG, and Cheng SJ.** Chromosome 3 imbalances are the most frequent aberration found in non-small cell lung carcinoma. *Lung Cancer* 23: 61–66, 1999.
7. **Mukasa A, Ueki K, Matsumoto S, Tsutsumi S, Nishikawa R, Fujimaki T, Asai A, Kirino T, and Aburatani H.** Distinction in gene expression profiles of oligodendrogliomas with and without allelic loss of 1p. *Oncogene* 21: 3961–3968, 2002.
8. **Pei J, Balsara BR, Li W, Litwin S, Gabrielson E, Feder M, Jen J, and Testa JR.** Genomic imbalances in human lung adenocarcinomas and squamous cell carcinomas. *Genes Chromosomes Cancer* 31: 282–287, 2001.
9. **Petersen S, Aninat-Meyer M, Schluns K, Gellert K, Dietel M, and Petersen I.** Chromosomal alterations in the clonal evolution to the metastatic stage of squamous cell carcinomas of the lung. *Br J Cancer* 82: 65–73, 2000.
10. **Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, and Brown PO.** Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* 23: 41–46, 1999.
11. **Reik W and Walter J.** Imprinting mechanisms in mammals. *Curr Opin Genet Dev* 8: 154–164, 1998.
12. **Virtaneva K, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri MA, Bloomfield CD, de La Chapelle A, and Krahe R.** Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc Natl Acad Sci USA* 98: 1124–1129, 2001.



Pure Appl. Chem., Vol. 75, Nos. 11–12, pp. 2405–2410, 2003.

Workshop 1.5

Fragment molecular orbital study of the binding energy of ligands to the estrogen receptor*

Kaori Fukuzawa[1], Kazuo Kitaura[2], Kotoko Nakata[3], Tsuguchika Kaminuma[4], and Tatsuya Nakano[3,‡]

[1]Fuji Research Institute Corporation, 2-3 Kanda Nishiki-cho, Chiyoda-ku, Tokyo 101-8443, Japan; [2]National Institute of Advanced Industrial Science and Technology, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan; [3]National Institute of Health Sciences, 1-18-1 Kamiyoga, Setagaya-ku, Tokyo 158-8501, Japan; [4]Chem-Bio Informatics Society, 4-3-16 Yoga #301, Setagaya-ku, Tokyo 158-0097, Japan

Abstract: We examined the published data for the binding affinity of typical ligands to the α-subtype of the human estrogen receptor with use of an approximate molecular orbital method applicable to interacting molecular clusters. An ab initio procedure for "molecular fragments" proposed recently to deal with such macromolecules as proteins was applied to the molecular orbital calculations. The receptor protein was primarily modeled using 50 amino acid residues surrounding the ligand. For a few ligand-receptor complexes, the binding energy was also calculated with use of 241 amino acid residues contained in the entire binding domain. No significant difference was found in the calculated binding energy between the complex modeled with ligand-surrounding 50 amino acids and that with residues of the entire domain. The calculated binding energy was correlated very well with the published relative binding affinity for typical ligands.

INTRODUCTION

The effect of estrogenic ligands is induced by their binding to the estrogen receptors (ERs) [1–3]. Since a variety of unknown compounds could bind to the ligand-binding domain (LBD) of the ER and exert hormone-like effects on human and wildlife health, the ER is an important research target for the development of therapeutic agents [3,4] as well as the screening of endocrine disruptors [5]. A number of experimental and theoretical efforts have been carried out for the mechanism of the interaction of ligands with the ER LBD. Most of the theoretical works, however, have stood on empirical force field approximations [6–8]. Although they are suited for calculating macromolecules in terms of the computational time, empirical approaches may not be accurate enough theoretically. Hoping to establish a time-saving and versatile computational procedure for biomacromolecules, we recently proposed the fragment molecular orbital (FMO) method [9]. Here, we report the result of our FMO study for the interaction of ligands with the α-subtype of ER carried out to elucidate its submolecular mechanism theoretically and accurately.

*Report from a SCOPE/IUPAC project: Implication of Endocrine Active Substances for Human and Wildlife (J. Miyamoto and J. Burger, editors). Other reports are published in this issue, *Pure Appl. Chem.* 75, 1617–2615 (2003).

‡Corresponding author

METHODS

In the FMO method [9], a single molecule or a molecular cluster (a group of molecules interacting to each other noncovalently) is dealt with after being divided into fragments to which electron pairs are assigned according to certain rules. The molecular orbitals (MOs) for fragments and fragment pairs (combinations of two fragments) are calculated under conditions under which the orbitals are forced to localize as the closed shell within the corresponding region. For fragments to which no electron pair is allocated from the bond when detached in the fragmentation, the MO is built from usual atomic basis functions of the constituent atoms according to the conventional linear combination of atomic orbitals to yield molecular orbitals (LCAO-MO) framework. For fragments in which bonding electron pair is left, the atomic valence basis function of the partner atom, with which the fragment is connected originally, is used additionally in the LCAO-MO model. The initial calculation for each fragment MO yields the initial electron density distribution.

The Hamiltonian for each fragment is composed to include the terms for the electrostatic potential governed by electrons in the surrounding fragments and all nuclei in the molecule. Since the electrostatic potential of each fragment depends on the electron distribution of surrounding fragments, the electron density distribution of each fragment is calculated first using the initial electron distribution calculated in a manner described above. A set of "Schrödinger" equations for every fragment with the initial electron density is solved iteratively until the electron density distribution for all fragments converges self-consistently. Likewise, the Hamiltonian of each fragment pair has the terms for the potential arising from electrons in the surrounding fragments and the terms from every nuclear charge in the molecule. The set of equations for fragment pairs is solved using the electrostatic potential from the converged electron density distribution of the surrounding fragments. The potential energy of fragments and fragment pairs at the HF/STO-3G level is calculated to estimate the energy of the total system.

The ligand molecules examined here are shown in Fig. 1. The coordinates of heavy atoms in the ER complex of EST, RAL, DES, and OHT were fixed as being equivalent to those of the PDB files, entries 3ERE, 1ERR, 3ERD, and 3ERT, of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB), respectively [10–12]. For ligands such as ESTA, GEN, TAM, BISA, BISF, CLO, and OHC, the PDB files for the ERα complex are not available. Thus, the binding geometry of the first two ligands was approximated first by superimposing the "phenoxy" substructure of the phenol moiety on that of EST in the 3ERE, while that of the others was by superimposing their "phenoxy" substructure or corresponding phenyl group on that of OHT in the 3ERT file. Then, the geometry of GEN was approximated by that in the ERβ-GEN complex taken from the PDB 1QKM file. TAM, CLO, and OHC were modeled with the Insight II system [13] based on the geometry of OHT, and the others were optimized using the HF/6-31G(d) method. The geometry of hydrogen atoms was modeled with the Insight II system [13] and the CHARMm force field calculations [14].

Hydrogen bonds, occurring between the ligand and surrounding residues directly as well as through the mediation of a single water molecule, have been shown to stabilize the ER ligand binding [15]. In this study, the most stable geometry of the hydrogen bond network was calculated at the HF/6-31G(d) level [16] with use of a model molecular cluster consisting of such hydrogen-bonding residues in the LBD as Glu 353, Leu 387, Arg 394, and His 524, each of the ligands and the single water molecule (Model 3).

The entire LBD of the receptor protein containing 241 amino acid residues (Model 1) was used for the calculation only for some ligands. The binding domain was, however, primarily modeled with use of 50 amino acid residues "directly" surrounding the ligand (Model 2) as displayed in Fig. 2. To make the fragmentation of the receptor protein, the peptide chain was divided at the Cα atom into blocks of every two residues in a manner as shown in Fig. 3. The ligand as well as the hydrogen-bonding water molecule was treated as a single fragment.

All the FMO calculations were carried out with an FMO program package, ABINIT-MP [17], mostly on dual Pentium III 1-GHz clusters equipped with 32 processor units. The time required for cal-

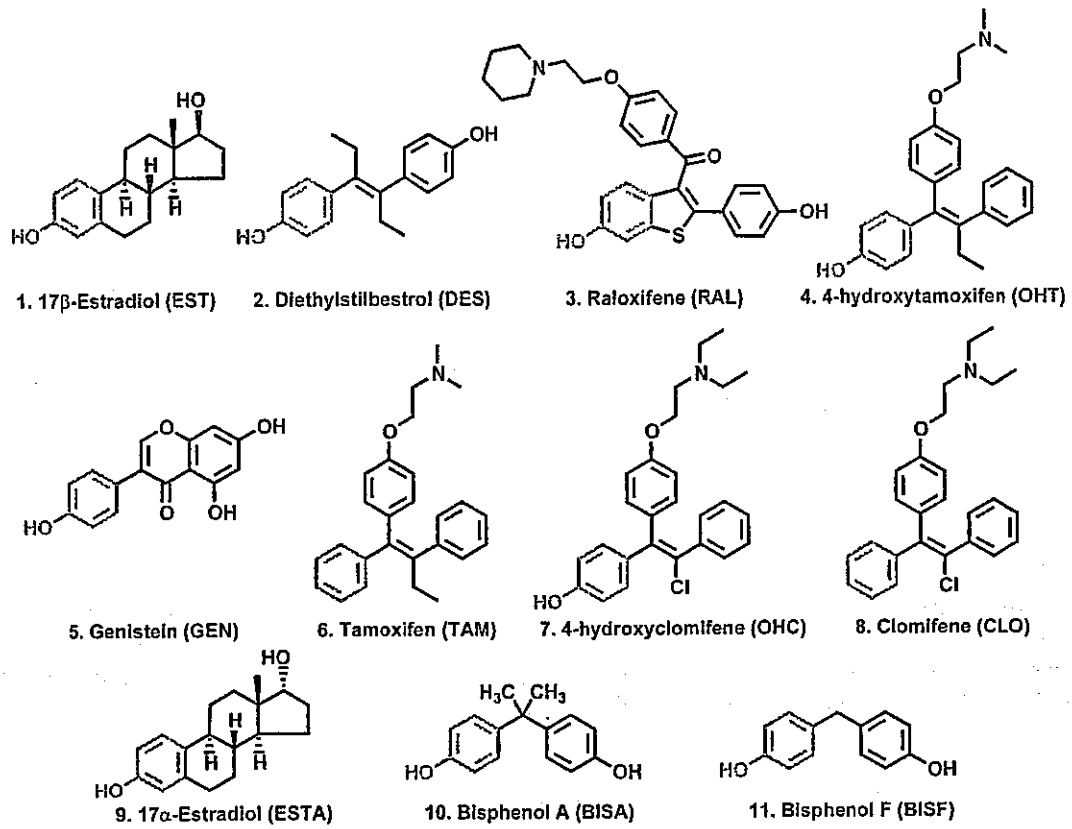


Fig. 1 Ligands used for the calculation of the binding energy. Light black substructures represent the moiety to be superimposed with the corresponding moiety in reference compounds.

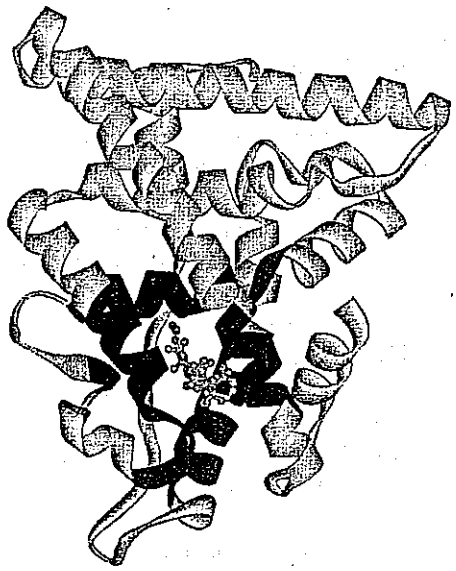


Fig. 2 The ribbon display of the ERα LBD complexed with 17β-estradiol (1, EST). Model 1 including 241 residues is shown as the entire picture. Fifty residues surrounding "directly" the ligand for Model 2 are dark-colored. The ligand and the water molecule are displayed inside the matrix using ball and stick.

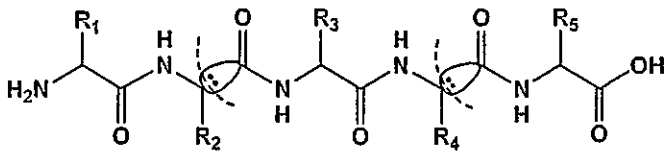


Fig. 3 Fragmentation of peptides indicated as broken arcs.

culating entire ERα LBD containing 241 residues with ca. 4000 atoms was about 14 h. The accuracy of the FMO method has been examined using crambin, a protein series with 46 residues [9]. The ab initio total energy values calculated at the HF/STO-3G level for [Pro^{22}, Leu^{25}]crambin with and without the FMO approximation are –17779.5030 and –17779.5024 a.u., respectively, corresponding to a difference below 0.5 kcal/mol. The computational time is "drastically" reduced with the FMO procedure compared to that without the FMO approximation.

RESULTS AND DISCUSSION

The energy of each of the three systems, i.e., the receptor, $E_{receptor}$, ligand E_{ligand}, and the ER ligand complex, $E_{complex}$, can be calculated from the sum of energy values of fragments and the counterpart for fragment pairs within each system under certain conditions [9]. In the calculation of the $E_{receptor}$ value, the hydrogen-bonding water molecule was included as a fragment along with "dipeptide" fragments. The binding energy for a given ligand (ΔE_{ligand}) can be expressed in eq. 1 as the difference in the energy between complex and components.

$$\Delta E_{ligand} = E_{complex} - (E_{receptor} + E_{ligand}) \quad (1)$$

The binding energy relative to that of 17β-estradiol (EST), $\Delta\Delta E_{ligand}$, in eq. 2 is the value to be compared with the experimental relative binding affinity (RBA) value. The RBA value of 17β-estradiol is defined as 100.

$$\Delta\Delta E_{ligand} = -(\Delta E_{ligand} - \Delta E_{EST}) \quad (2)$$

The $\Delta\Delta E_{ligand}$ values estimated using Model 2 are plotted against the published values of log (RBA/100) in Fig. 4.

The ligands 1–6, 9, and 10, of which the experimental RBA value is known, are shown as a circle in Fig. 4. For these 8 compounds, the correlation between $\Delta\Delta E$ and log (RBA/100) seems to be promising, the correlation coefficient r being 0.837. In particular, there is a very good correlation (r = 0.931) for the 7 ligands omitting TAM (6). From the correlation equation (n = 8), the log (RBA/100) value of ligands 7, 8, and 11, of which the RBA value is unknown, can be estimated with use of the calculated $\Delta\Delta E$ value. These 3 compounds are shown as a square in the plot.

The $\Delta\Delta E$ value was also calculated according to Model 1 for the complex of ligands 1~4. The result was almost identical with that calculated with Model 2. The difference in the $\Delta\Delta E$ value between two models was mostly below 3 kcal/mol, suggesting that the binding between ER and ligand is local. Another interesting finding was a difference in the charge distribution between complexed and individual component molecules. The total charge of ligands was changed to be negative with the values –0.00 ~ –0.18 when complexed with ER. The greatest negative charge influx occurs from Glu 353 to ligands, and a slight efflux is observed into Arg 394 and His 524. Such charge transfer is highly related with the binding energy. In fact, the ΔE tends to be greater with the increase in the difference of the charge distribution. Thus, most of the stabilization in the ER–ligand docking arises from the ligand-Glu 353 interaction. This observation seems to indicate that the charge is variable in the ER–ligand interaction, and therefore atomic charges should be calculated dynamically instead of using fixed charges as in classical calculations.

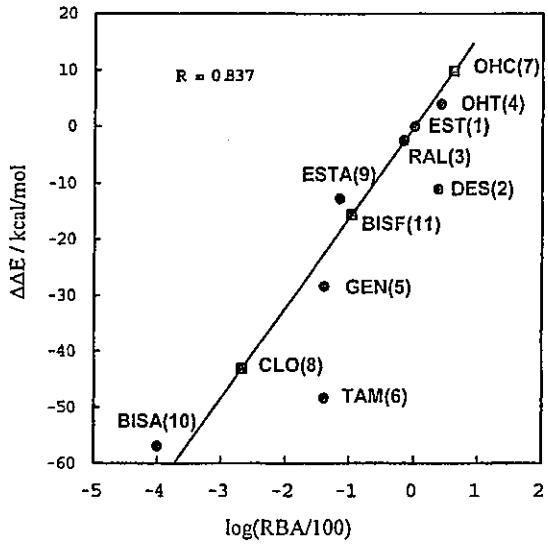


Fig. 4 Relationship between calculated relative binding energy ($\Delta\Delta E$) and experimental relative binding affinity [log (RBA/100)] of eight ligands (●), and the estimation of log (RBA/100) for three ligands (■). The regression line is drawn so that it is forced to pass the origin of coordinates.

To summarize, we have applied the ab initio FMO method to ER ligand binding which allows us to accurately predict the relative binding energy of xenoestrogenic ligand molecules from a "single" energy calculation. Given a variety of compounds, some of which could bind to the ER, such methods as we have proposed may provide a powerful tool for assessing the affinity of putative xenoestrogens in silico prior to biological studies. For further improvements, it is necessary to optimize not only the hydrogen bond, but also the geometry of the ligand and surrounding residues to estimate possible effects, in particular, those according to induced-fit in the ER ligand binding. Such functions are under development in our group.

REFERENCES

1. K. Paech, P. Webb, G. G. J. M Kuiper, S. Nilsson, J.-Å. Gustafsson, P. J. Kushner, T. S. Scanlan. *Science* **277**, 1508–1510 (1997).
2. G. G. J. M. Kuiper, J. G. Lemmen, B. Carlsson, J. C. Corton, S. H. Safe, P. T. van der Saag, B. van der Burg, J.-Å. Gustafsson. *Endocrinology* **139**, 4252–4263 (1998).
3. T. Barkhem, B. Carlsson, Y. Nilsson, E. Enmark, J.-Å. Gustafsson S. Nilsson. *Mol. Pharmacol.* **54**, 105–112 (1998).
4. S. Nilsson, G. Kuiper, J.-Å. Gustafsson. *Trends Endocrinol. Metab.* **9**, 387–395 (1998).
5. C. Sonnenschein and A. M. J. Soto. *J. Steroid Biochem. Molec. Biol.* **65**, 143–150 (1998).
6. S. P. Bradbury, O. G. Mekenyan, G. T. Ankley. *Environ. Toxicol. Chem.* **17**, 15–25 (1998).
7. B. C. Oostenbrink, J. W. Pitera, M. M. H. van Lipzig, J. H. N. Meerman, W. F. van Gunsteren. *J. Med. Chem.* **43**, 4594–4605 (2000).
8. P. D. Kirchhoff, R. Brown, S. Kahn, M. Waldman, C. M. Venkatachalam. *J. Comput. Chem.* **22**, 993–1003 (2001).
9. T. Nakano, T. Kaminuma, T. Sato, K. Fukuzawa, Y. Akiyama, M. Uebayasi, K. Kitaura. *Chem. Phys. Lett.* **351**, 475–480 (2002).
10. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne. *Nucleic Acids Res.* **28**, 235–242 (2000); <http://www.rcsb.org/pdb/>.

11. A. M. Brzozowski, A. C. W. Pike, Z. Dauter, R. E. Hubbard, T. Bonn, O. Engström, L. Öhman, G. L. Greene, J.-Å. Gustafsson, M. Carlquist. *Nature* **389**, 753–758 (1997).
12. A. K. Shiau, D. Barstad, P. M. Loria, L. Cheng, P. J. Kushner, D. A. Agard, G. L. Greene. *Cell* **95**, 927–937 (1998).
13. InsightII Version 98.0, Molecular Simulations Inc., San Diego, CA (1998).
14. CHARMm, Version 25.2, Revision: 98.0731.
15. D. M. Tanenbaum, Y. Wang, S. P. Williams, P. B. Sigler. *Proc. Nat. Acad. Sci. USA* **95**, 5998–6003 (1998).
16. Gaussian 98, Revision A.7, M. J. Frisch et.al., Gaussian, Inc., Pittsburgh PA (1998).
17. ABINIT-MP: <http://moldb.nihs.go.jp/abinitmp/>.

ANALYTICAL SCIENCES NOVEMBER 2003, VOL. 19

Test of Significant Differences with *a priori* Probability in Microarray Experiments

Kyoko TODA,† Seiichi ISHIDA,† Kotoko NAKATA, Rieko MATSUDA, Yukari SHIGEMOTO-MOGAMI, Kayoko FUJISHITA, Shogo OZAWA, Jun-ichi SAWADA, Kazuhide INOUE, Koichi SHUDO, and Yuzuru HAYASHI††

National Institute of Health Sciences, 1-18-1 Kami-Yoga, Setagaya, Tokyo 158-8501, Japan

A traditional method for comparing two expression levels of genes in microarray experiments is the two-sample *t*-test. Because of the difficulty in using a large number of microarrays, an alternative method is required which can provide a reliable judgment of the comparison from a small number of replicates, even from a single pair of control and treatment. We present a method for detecting the changes in the gene expression levels under two different conditions in microarray experiments. Our method targets a single experiment for each condition, while retaining the statistical advantages of the *t*-test. The new proposals are: 1) standard deviation (SD) estimates of the expression levels which are an indicator for significant differences are given *a priori* as a function of the expression levels; 2) the limit of detection (LOD) for the expression levels is used to eliminate the majority of genes expressed at extremely low levels. The *a priori* SD estimates are obtained from six replicates under a fixed condition and are shown to be the approximate, but proper description of the expression uncertainty covering diverse conditions (*e.g.*, different samples (human and rat) and different DNA chips). The LOD is defined as three times blank SD according to the IUPAC recommendation. A cell line (HL60) which will undergo macrophage differentiation on treatment with 12-*O*-tetradecanoylphorbol 13-acetate (TPA) is taken as an example. Our method is compared with the *t*-test for the data on duplicate TPA experiments and the former alone is evaluated with the data on a single TPA experiment. The errors from sample preparation and instrumental analysis are discussed.

(Received June 6, 2003; Accepted September 16, 2003)

Introduction

The advent of high-throughput array technology has now made it possible to collect data on thousands to tens of thousands of genes simultaneously. However, methods for detecting the genuine changes in the gene expression levels in cells or tissues are still evolving.[1-11]

A straightforward method for comparing two expression levels of genes is the traditional two-sample *t*-test. The basic problem with the *t*-test in microarray experiments, however, is that the repetition is restricted within a small number in most cases, because experiments are costly or tedious to repeat. Although the importance in replication has been illustrated,[1,3,6,9] situations often arise where only single or duplicate experiments for each condition are allowed.

The purpose of this paper is to put forward a method for testing the significant differences of the gene expression levels under a single pair of experiments (a control experiment and treatment experiment). In order to take into account the stochastic aspects of gene expressions, we model our algorithm on the *t*-test. In our approach and the *t*-test, the SD estimates of the gene expression levels are a criterion for the statistical judgment, but one of the key differences is how to estimate the SD for each gene.

In the *t*-test, the SD estimates are derived from the same data set as those to be judged by the *t*-test, itself. This fact can be closely connected with the above-mentioned problem of replication. In our approach, the SD estimates, referred to here as *a priori* SD, are obtained from experimental results which are different from the target data set of the judgment.

Statistics tells that the variability in the estimates of SD obeys the chi-squares distribution and is much larger than the variability in the estimates of averages, as long as the estimates are obtained by repetition. That is, the estimates of averages are more reliable. The *t*-test uses the SD estimates directly, but in this paper, the *a prior* SD is given as a function of the average of the gene expression levels. Then, we can easily expect that our approach can provide more stable judgment, but needs a sound model for the *a prior* SD.

The idea of the *a priori* SD is not novel in the area of analytical chemistry. Since more than three decades ago, there have been published many theories and methods for estimating SD with no recourse to repetition, especially in instrumental analyses.[12-27] In spite of varied symbols and terminology in the literature, the largest part of uncertainty equations proposed can take a universal form:[24]

$$\mathrm{RSD}^2 = \frac{s_B^2}{A^2} + I^2, \quad (1)$$

where RSD denotes the relative standard deviation of measurements, s_B denotes blank SD, A measurements (*e.g.*, area), and I independent error. To our knowledge, Huber *et al.*

† Co-first authors.

†† To whom correspondence should be addressed.

K. S. present address: Japan Pharmaceutical Information Center.

first used Eq. (1) in 1971. The identical error models were adopted in microarray experiments.[4,6] This paper also follows suit.

The mathematical formalism of uncertainty like Eq. (1) has wide applicability. Examples are LOD,[21,24,26,28] confidence intervals of linear calibration[28] and tests of significant differences.[3,4,6] Theoretical SD descriptions elaborated so far include Winefordner's theory,[18-20] Ingle's theory,[15-17] Bouman's theory[21,22] and FUMI theory (FUnction of Mutual Information).[23-25]

Proceeding along the lines suggested by the FUMI theory, our approach is named after it. The other salient feature of the FUMI theory in this paper is the introduction of LOD which is helpful to remove the vast majority of genes expressed at exceedingly low levels. This point is also a problem with which simple fold-change methods are accompanied.[3,8] The FUMI theory is applied to a cell line (HL60) which will undergo macrophage differentiation on exposure to a tumor promotor, 12-*O*-tetradecanoylphorbol 13-acetate (TPA).[29,30]

The errors due to sample preparation, before the instrumental measurement, are often a critical problem in practice. Typical experiments are planned to discern the contributions of the preparation and measurement processes to the total analytical error. This paper demonstrates that the error magnitude of preparation is even smaller than that of the measurement in our analytical system.

Materials and Methods

There are about ten thousand genes on a DNA chip used (GeneChip, Affymetrix). A probe set for a given gene on the DNA chip usually contains sixteen probe pairs, each of which is made up with perfect match and mismatch probe cells. The total RNAs prepared from a sample are enzymatically converted into fragmented, biotin-labeled cRNAs and hybridized to the probe sets. After washing and staining with phycoerythin conjugated streptavidin, the amount of hybridized cRNAs is quantified by scanning the DNA chip with the argon-ion laser scanner. The resulting fluorescence image data are processed and given as "Signal" by a software (Microarray Suite 5.0, Affymetrix). The values of "Signal" can directly be related with the expression levels of the genes and are used as measurements of samples by the *t*-test and FUMI theory.

All the experiments including RNA isolation, hybridization, *etc.* were carried out according to the manufacturer's protocol. The cell lines used were human hepatocellular carcinoma cell line (HepG2), human promyelocytic leukemia cell line (HL60) and rat microglia. The rat cell line was obtained from primary cell cultures of neonatal Wister rat brains as described previously.[31] The combinations with DNA chips (GeneChip, Affymetrix) were: HepG2 (U95A); HepG2 (U95B); HL60 (U95A); rat microglia (U74A).

The biotin-labeled cRNA for each cell line was stocked and later used for the repeated experiments (n = 6) which began with the hybridization (six arrays for each cell line).

The HL60 cells were exposed to 20 nM TPA for 1 h and the biotin-labeled cRNA was prepared and stored as a stock solution. A total of four U95A arrays were used (two with the TPA-exposed stock solution and two with the control stock solution).

The model experiments for the entire microarray analyses were carried out as follows: the total RNA was prepared from eleven culture dishes of HL60 by the RNeasy Mini total RNA preparation kit (Qiagen, Germany); cRNA was synthesized from 10 μg of total RNA on each dish according to the Affymetrix protocol; the cRNA was determined by ultra-violet absorption spectrometry; the RSD was calculated from the measurements (n = 11).

Theory

A brief review of the *t*-test and an in-depth explanation of our test are given below.

t-test and Cochran-Cox method

It is assumed that the number of replicates is two for exposure and control experiments, respectively. Let $\bar{X}_E$ be the mean of the expression levels (measurements), X_E, of a gene for exposed samples and $\bar{X}_C$ be the mean of measurements, X_C, of the gene for control samples. In the *t*-test, the expression levels of a gene are judged to be significantly different, if the absolute difference between $\bar{X}_E$ and $\bar{X}_C$, meets the condition:[32]

$$\frac{|\bar{X}_E - \bar{X}_C|}{s} > 9.925, \tag{2}$$

where 9.925 is the critical value of |t| at a significant level of 1%. Here, the SD estimate, s, takes the form:[32]

$$s = \sqrt{(s_E^2 + s_C^2)/2}\,, \tag{3}$$

where s_E and s_C are the SD estimates of individual measurements, X_E and X_c, of exposure and control, respectively.

Before the *t*-test, the *F*-test is carried out for the SD estimates, s_E and s_C. If the homoscedasticity assumption ($s_E = s_C$) is rejected by the *F*-test, another critical value, 63.657, is used in Eq. (2) instead of 9.925 (Cochran-Cox method).[33] If the homoscedasticity is accepted, the *t*-test (Eqs. (2) and (3)) follows.

FUMI theory

The FUMI theory has an equivalent formalism of judgment:

$$\frac{|\bar{X}_E - \bar{X}_C|}{s} > 2.58 \tag{4}$$

where s means the SD estimate of numerator, $\bar{X}_E - \bar{X}_C$, and 2.58 is the critical value for a significant level of 1% under the assumption that the distribution of $\bar{X}_E - \bar{X}_C$ is normal.

We define σ (*a priori* SD) as the SD of the individual measurements, X_E and X_C (see Eq. (6)). If the SD of $\bar{X}_E$ is equal to the SD of $\bar{X}_C$, then the SD, s, of $\bar{X}_E - \bar{X}_C$ can be given:

$$s = \sigma \qquad (= \sqrt{2}\,(\sigma \div \sqrt{2}\,);\ \text{note that } n = 2). \tag{5}$$

Figure 1 illustrates the algorithm of the FUMI theory which consists of three types of judgment based on the *a priori* SD, σ (gray rhombi).

Step 1 (Difference > 3s; $s = \sqrt{2}\,\sigma$): If the duplicate expression levels for a gene under the same conditions (*e.g.*, X_C for Control 1 and Control 2) are even more different than those expected by the *a priori* SD, the gene is eliminated from the analysis.

Step 2 (Difference > 2.58s; $s = \sigma$; see Eqs. (4) and (5): If the difference in the mean expression levels ($= \bar{X}_E - \bar{X}_C$) is regarded as being significant at 1% level, the gene goes to the next step. If not, it is discarded.

Step 3 (Larger data > 2 × LOD; LOD = $3(\sigma \div \sqrt{2}\,)$): If the largest level of the control and exposure means is less than

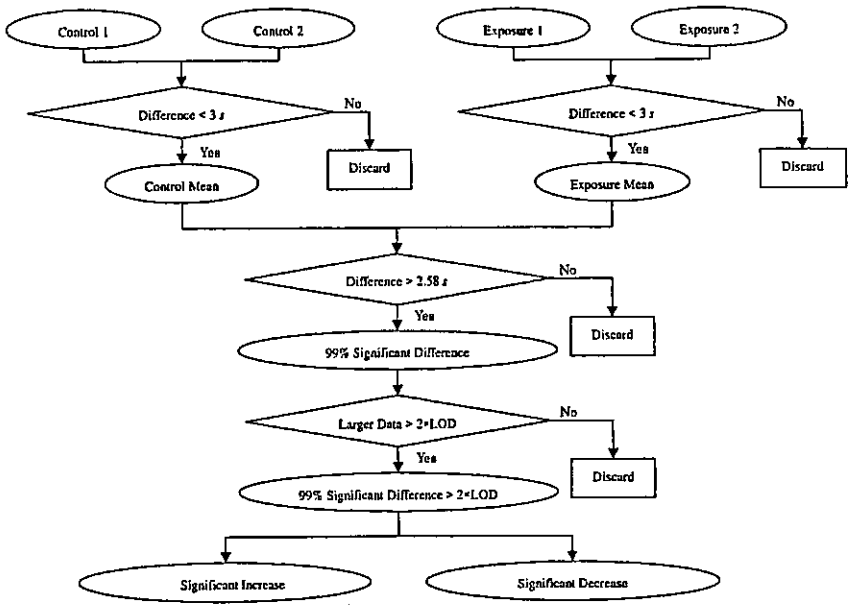


Fig. 1 Flow chart of probabilistic significance test (FUMI theory). *s*, SD; LOD, limit of detection.

twice the limit of detection (LOD), the gene is removed from the analysis.

According to the IUPAC recommendation,[34] LOD is defined as three times blank SD, s_0 ($s_0 = \sigma$ where $X = 0$ in Eq. (6)). The blank measurements correspond to the expression levels of house-keeping genes and are assumed to vary due to a random fluctuation of detector noise or other error sources. The above LOD definition implies that the probability for a noise-created false signal being above the LOD is at most 0.13%.[34] If the smallest level of the control and exposure means is just the LOD, the minimum level which can be distinguished from the LOD level at a significant level of 0.13% is $2 \times$ LOD.

To a single pair of experiments, the FUMI theory can also be applied. However, Step 1 should be skipped and the test begins at Step 2. Moreover, the critical values should be changed: $s = \sqrt{2}\sigma$ in Step 2; LOD = 3σ in Step 3.

Results and Discussion

Precision of microarray measurement

Figure 2A shows the precision plot for human HepG2 using U95A chips. The X axis is the average of 6 expression levels (measurements) for each gene (total 12559 genes). The Y axis denotes the SD values estimated statistically from the 6 measurements each. The SD estimates (•) are not randomly scattered, but seem to increase with increasing expression level. This trend of the precision plot is quite common to many instrumental analyses such as ultraviolet-visible absorption spectrometry, atomic absorption spectrometry and high performance liquid chromatography.[15,23,25] The similar precision plots for microarrays were observed.[3,4,6]

Our microarray experiments were repeated over a part of the entire process, ranging from the hybridization on the chips to the data processing which gives the measurements, X. The least squares fitting to the observed SD values in Fig. 2A can lead to the SD dependence on X (for details, see the legend of Fig. 2):

$$\sigma = \sqrt{0.009639\,X^2 + 91897.8} \qquad (6)$$

This is the *a priori* SD defined in the preceding section and is shown in Fig. 2A (——). As for Fig. 2A, exceptionally large SD estimates are spotted frequently at high expression levels and the region of the least squares fitting is limited as described in the figure legend to guarantee the goodness of fit. Li and Wong revealed outliers due to various reasons including image artifacts in oligonucleotide microarrays.[10]

The dots (•) of the precision plots in Figs. 2B - D are the SD estimates observed under the conditions different from Fig. 2A (*i.e.*, different samples, chips; see the legend). However, the lines (——) in Figs. 2B - D are just the *a priori* SD, σ, drawn in Fig. 2A (Eq. (6)). From this fact, we can see that although the *a priori* SD is phenomenological without knowledge about the causality of errors appearing on X, the *a priori* SD can provide a general aspect in the microarray experiments conducted here. The results of the TPA experiments are analyzed below with Eq. (6).

The distribution of microarray measurements is shown in Fig. 3A. The data are collected from the genes which give almost the same averages, X_D (= 1100), of measurements (for details, see the legend). Figure 3B illustrates the normal distribution with the SD obtained by substituting X_D for X in Eq. (6). Although there are slight differences between the observed and normal distributions, especially around the center and on the edges, the measurements can be considered normally distributed. This normality makes the SD scattering pattern in Fig. 2 interpretable in terms of the chi-square distribution. A log-normal distribution was observed for the bulk of Affymetrix microarray spot intensities.[35]

Underlying the *t*-test is the gene-specific SD, or rather changeable SD from gene to gene. However, the intensity-dependent SD of the FUMI theory (Eq. (6)) is not surprising in the field of instrumental analysis. A photomultiplier can know the intensity of light, but can never know anything else, *e.g.*, the origin of light is a human gene or rat gene.

Error sources and evaluation of a priori SD

The error sources of the microarray measurements have yet to be identified. The constant term of Eq. (6) (= 91897.8)

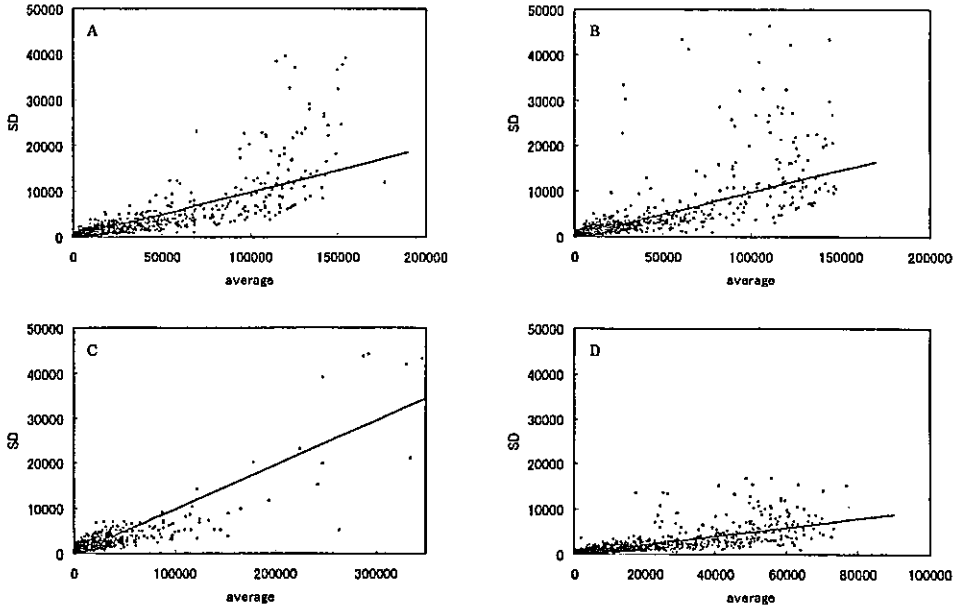


Fig. 2 Precision plots for microarray measurements in different conditions. Six DNA chips are used for each condition. The average and SD estimate of 6 measurements for a gene leads to the values of X and Y axes, respectively. •, the SD estimates from six replicates; —, the fitted line (*a priori* SD, Eq. (6)). Conditions (samples, chips): A, human HepG2, U95A; B, human HL60, U95A; C, human HepG2, U95B; D, rat microglia, U74A. The intercept (= 91897.1) of Eq. (6) is the average of the variance estimates over X from 1000 to 2000. The coefficient (= 0.009639) is obtained from the least squares fitting of a straight line passing the origin to the average-subtracted variance estimates over X from 1000 to 100000.

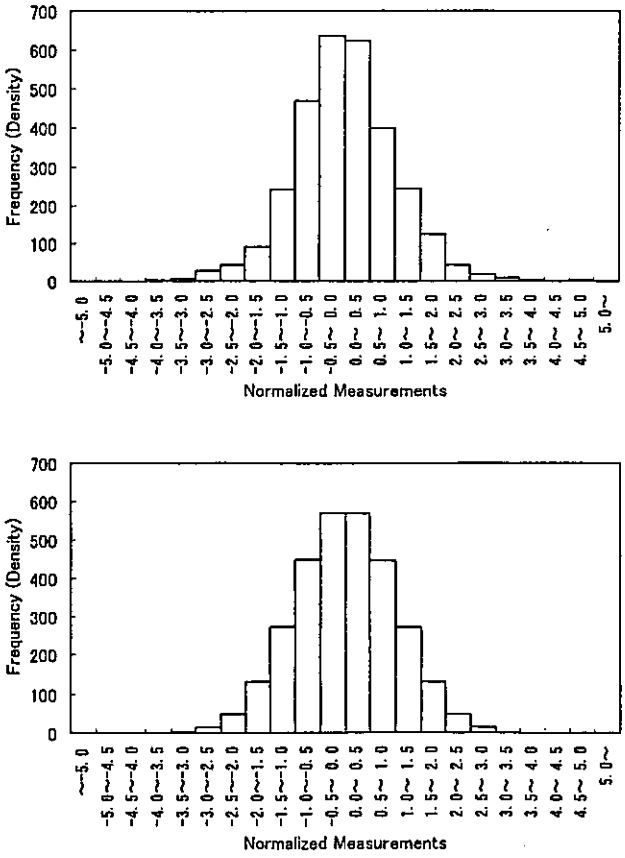


Fig. 3 Distribution of expression levels (top) and normal distribution (bottom). The top figure uses 498 genes which are located in the X region from 1000 to 1200. The total number of measurements used is 2988 (= 6 × 498).

dominates at low fluorescence intensities and will correspond to the background noise which comes mainly from the photomultiplier of the detection unit. The coefficient of X^2 (= 0.009639) plays an important role at high intensities where the RSD of measurements is almost invariant (~10% here). Promising candidates of error sources are some procedures before the light detection such as the incorporation of fluorescent tags and hybridization.

The model experiments, using no microarrays (see "Materials and Methods"), include the former part of typical entire analysis (preparation of total RNA and synthesis of cRNA), but the procedure corresponding to detection is quite simple (ultraviolet-visible absorption). The RSD for the model experiments was observed to be about 5%. This result implies that the experimental error originates mainly from the total RNA preparation and cRNA synthesis, since the precision of the UV detection is usually high (RSD < 1%).

Our microarray experiments lack the former part of the entire analysis (total RNA preparation and cRNA synthesis). However, even if they included it, the error of the entire analysis (RSD = $(10^2 + 5^2)^{1/2}$ = 11.2) would be almost equal to the error of the experiments without the former part (RSD ~10%, see above).

From the above discussion, it follows that in our microarray experiments, the most important error sources are the background noise at low fluorescence intensities and the incorporation of fluorescent tags and hybridization at high intensities. Since the former part of analysis does not affect the precision substantially, the *a priori* SD (Eq. (6)) can be considered to be applicable to the usual microarray experiments including the sample preparation.

t-test and FUMI theory for duplicate pairs of TPA experiments

If the sample size is large, the *t*-test and FUMI theory would