IPSJ Transactions on Bioinformatics Vol. 3 70–81 (Oct. 2010)

Original Paper

Differentially Aberrant Region Detection in Array CGH Data Based on Nearest Neighbor Classification Performance

Yuta Ishikawa[†1] and Ichiro Takeuchi[†1]

Array CGH is a useful technology for detecting copy number aberrations in genome-wide scale. We study the problem of detecting differentially aberrant genomic regions in two or more groups of CGH arrays and estimating the statistical significance of those regions. An important property of array CGH data is that there are spatial correlations among probes, and we need to take this fact into consideration when we develop a computational algorithm for array CGH data analysis. In this paper we first discuss three difficult issues underlying this problem, and then introduce nearest-neighbor multivariate test in order to alleviate these difficulties. Our proposed approach has three advantages. First, it can incorporate the spatial correlation among probes. Second, genomic regions with different sizes can be analyzed in a common ground. And finally, the computational cost can be considerably reduced with the use of a simple trick. We demonstrate the effectiveness of our approach through an application to previously published array CGH data set on 75 malignant lymphoma patients.

1. Introduction

Array Comparative Genomic Hybridization (array CGH) is a useful technology for measuring DNA copy numbers in genome-wide scale. In an array CGH analysis of a tumor cell, the tumor DNA and a normal (reference) DNA are cohybridized to a microarray of thousands of genomic clones of BAC, cDNA, or oligonucleotide probes [1)]. For each of thousands of probes, an array CGH experiment returns the $\log_2$-ratio of the number of DNA copies in the tumor cell to that in the normal (reference) cell at the genomic region corresponding to the probe. A $\log_2$-ratio greater (less) than zero indicates a possible gain or amplification (loss or deletion) in DNA copies in the tumor cell at the genomic region.

†1 Nagoya Institute of Technology

Figure 1 shows three examples of array CGH data obtained by array CGH analyses of three tumor cells taken from three different lymphoma patients [2)]. In the figure, the numbers $1, \cdots, 22$ in the horizontal axis indicate chromosome 1, $\cdots$, chromosome 22, respectively[⋆1], and the measurements in the vertical axis represent the $\log_2$-ratios at each of 2,035 BAC probes.

The first fundamental task in array CGH data analysis is finding aberrant (amplified or deleted) genomic regions in a single array (e.g., for an individual patient). Many computational algorithms have been developed for this task [3)–5)]. The next level of task is identifying common aberrant regions in a group of arrays (e.g., for a group of patients in common clinical condition). Compared to the first task, the smaller number of studies has been done for this task [5)]. However, such common aberrant regions are usually of great interest in many biological studies. Currently, in these studies, common aberrant regions are rather informally defined. One such informal approach is to simply compute the average $\log_2$-ratios of the group of arrays and to apply some algorithms for the first task (for detecting aberrant regions in a single array) to the average $\log_2$-ratios. The third level of task in array CGH data analysis is detecting differentially aberrant regions in two or many groups of arrays, i.e., detecting the regions that are commonly amplified or deleted in one group of arrays, but not in the other group(s) of arrays. As long as we know, there is no formal treatment in this task in the literature, although such differentially aberrant regions can often provide important implications in biological researches. More rigorous and formal approaches are needed for the second and the third tasks.

In developing computational algorithms for array CGH data analysis (all in the above three tasks), it is important to take into account the spatial relationship of probes. Unlike gene expression microarray (which measures mRNA expression level of each gene), a probe in CGH microarray represents a segment of genomic DNA, and thus they are considered to be sequentially connected in each chromosome. For example, we can observe many contiguously amplified or deleted regions in Fig. 1. For this reason, most of the above mentioned algorithms for the

⋆1 In most array CGH analyses (our experiment in Section 4 as well), non-sex 22 chromosomes are analyzed.

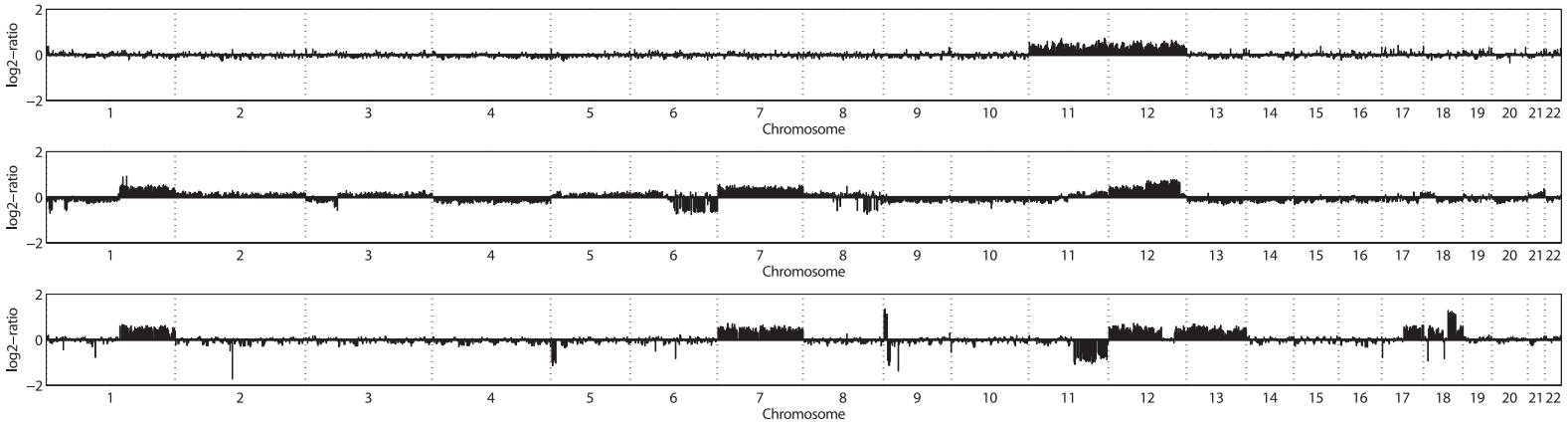


Fig. 1 Examples of array CGH data taken from three lymphoma patients: The numbers in the horizontal axis denote the chromosome and the measurements in the vertical axis indicate $\log_2$-ratio of each probe. See Section 4 for the detail of these arrays.

first task [3)–5)] were developed by translating the aberrant region detection problem into segmentation or breakpoint identification problem of a sequential data. On the other hand, for the second and the third tasks, it is difficult to incorporate spatial correlation among probes because we have to deal with multi-dimensional sequential data.

In the third task of detecting differentially aberrant regions, it is informative if we could certify the statistical significance of the "difference". In expression microarray data analysis, statistical significance analysis and related multiple comparison issue have been intensively studied [6),7)]. It is possible to apply those methodologies also to array CGH analysis if we ignore the spatial relationship among probes and we are just interested in the difference in genomic regions represented by each single probe. However, if we consider the spatial factor and interested in regions consisting of multiple consecutive probes, the problem of assessing their statistical significance comes with several difficulties.

At first, for regions consisting of multiple probes, we need to introduce *multivariate test* rather than univariate test (such as t-test, U-test, etc.) for computing the statistical significance in the region. Secondly, if we consider all possible consecutive regions in each of 22 chromosomes, the number of candidate regions would be quite large. For example, the BAC array in Fig. 1 consists of 2,035 probes, and the number of possible candidate regions would be 141,452. It means that in multiple-testing adjustment we need to work with huge amount of multiplicity of highly correlated test statistics (correlated because we consider many overlapped regions). Finally, we need to introduce a multivariate statistic that is independent of the dimensions of the data because we have to compare, for example, a region corresponding to a single probe and a region consisting of entire chromosome on an equal footing.

In this paper we propose an approach for the third task of detecting differentially aberrant regions using nearest neighbor (NN) multivariate test. NN multivariate test is a simple but powerful multi-sample multivariate test, in which some measures of nearest neighbor coincidence is used as a test statistic [8),9)]. Although there are many ways to quantify nearest neighbor coincidence, we use leave-one-out cross-validation (LOOCV) error of NN classifier as our test statistic$^{\star 1}$. To understand why LOOCV error of NN classifier can be used as a multivariate test

$\star$1 We can use LOOCV (or any other generalization error estimates) of other classification algorithms (such as SVM or decision tree) as a multivariate test statistic as long as we can compute the null distribution.

statistic, consider, for example, a binary classification problem for a multivariate two samples. If these two samples are differentially distributed, LOOCV error of NN classifier would be small, while if these two samples are identically distributed, LOOCV error of NN classifier would be large. It suggests that LOOCV error of NN classifier can quantify, in some sense, the difference between two or more multivariate samples.

In the remainder of the paper, we will show that the aforementioned three difficulties can be addressed (at least alleviated) with the use of NN multivariate test. To demonstrate the effectiveness of our approach, we analyze BAC array CGH data for 75 malignant lymphoma patients [2),10),11)]. The rest of the paper is organized as follows. In the next section, we formulate the problem of detecting differentially aberrant regions in two/multi-samples of CGH arrays, and discuss the aforementioned three difficult issues in more detail. In Section 3, we introduce NN multivariate test and clarify how it can address these three difficult issues. In Section 4, we analyze previously published array CGH data of 75 malignant lymphoma patients and identify statistically significant differentially aberrant regions. In addition, we describe an attempt to use the detected regions for tumor classification task, in which the goal is to classify two lymphoma subtypes based on array CGH data. Finally, we conclude this paper in Section 5.

2. Problem Formulation

In this section we first formulate the problem of detecting differentially aberrant region in two groups of CGH arrays. In order to simplify the description, we restrict our attention to two-sample problems, but it can be straightforwardly extended to multi-sample problems. After the problem formulation we discuss the three difficult issues underlying the problem.

2.1 Formulation

Suppose we have N CGH arrays and they are from two distinct groups (or classes in binary classification terminology) C_1 and C_2. Let N_1 and N_2 be the number of arrays in C_1 and C_2, respectively (i.e., $N_1 + N_2 = N$). Furthermore, the number of probes in c-th chromosome is denoted as ℓ_c for chromosomes $c = 1, \cdots, 22$, and each probe in chromosome c is indexed by $1, \cdots, \ell_c$. The $\log_2$-ratio at the probe j in chromosome c of array i is denoted as x_{icj}, for $i, = 1, \cdots, N$, $c, = 1, \cdots, 22$, and $j = 1, \cdots, \ell_c$. The class label of array i is represented as y_i: if array i is in C_1, $y_i = 1$, while if it is in C_2, $y_i = 2$.

To simplify the discussion, let us consider for a moment that the task of detecting aberrant regions in chromosome c of array i (i.e., the first task in the previous section for chromosome c of array i), for which we have ℓ_c $\log_2$-ratio signals $x_{ic1}, x_{ic2}, \cdots, x_{ic\ell_c}$. As noted in the previous section, probes in CGH arrays have spatial relationship because each probe is corresponding to physical genomic DNA region. In aberrant region detection task we need to investigate all possible sub-sequence of $x_{ic1}, x_{ic2}, \cdots, x_{ic\ell_c}$. The total number of sub-sequences is $\sum_{h=1}^{\ell_c} h = \frac{1}{2}\ell_c(\ell_c + 1)$ because we have one sequence with length ℓ_c (corresponding to the entire chromosome), two sub-sequences with length $\ell_c - 1$, $\cdots$, and, ℓ_c sub-sequences with length 1 (corresponding to single probe). Hereafter, we denote the number of all possible regions in the entire genome as $M \equiv \sum_{c=1}^{22}\sum_{h=1}^{\ell_c} h = \frac{1}{2}\sum_{c=1}^{22}\ell_c(\ell_c + 1)$, and each region is indexed by $m = 1, \cdots, M$. In addition, the set of index pairs (c, j) in region m is denoted as $\mathcal{R}_m$, and the number of probes in region m is written as $|\mathcal{R}_m|$. In the case of BAC arrays in Fig. 1, $M = 141,452$.

It would probably help the reader's understanding if we summarize the data in our task by an N-by-M input "matrix"

$$X = \begin{bmatrix} \{x_{1cj}\}_{(c,j)\in\mathcal{R}_1} & \{x_{1cj}\}_{(c,j)\in\mathcal{R}_2} & \cdots & \{x_{1cj}\}_{(c,j)\in\mathcal{R}_M} \\ \{x_{2cj}\}_{(c,j)\in\mathcal{R}_1} & \{x_{2cj}\}_{(c,j)\in\mathcal{R}_2} & \cdots & \{x_{2cj}\}_{(c,j)\in\mathcal{R}_M} \\ \vdots & \vdots & \ddots & \vdots \\ \{x_{Ncj}\}_{(c,j)\in\mathcal{R}_1} & \{x_{Ncj}\}_{(c,j)\in\mathcal{R}_2} & \cdots & \{x_{Ncj}\}_{(c,j)\in\mathcal{R}_M} \end{bmatrix},$$

and N-vector of labels $\boldsymbol{y} = [y_1, y_2, \cdots, y_N]^\top$, where, note that (i, m)-th entry of the input "matrix" X itself is defined by $|\mathcal{R}_m|$ real-valued vector for all $(i, m) \in \{1, \cdots, N\} \times \{1, \cdots, M\}$.

Figure 2 shows some examples of $\log_2$-ratio sequences in chromosome 1. In this CGH array, there are 173 BAC probes in chromosome 1 (i.e., $\ell_1 = 173$). The figure illustrates the aforementioned candidate regions. We consider all possible subsequences of the entire sequence of the 173 probes. The number of possible candidate regions in this example is $\frac{1}{2} \times 173 \times 174 = 15,051$.

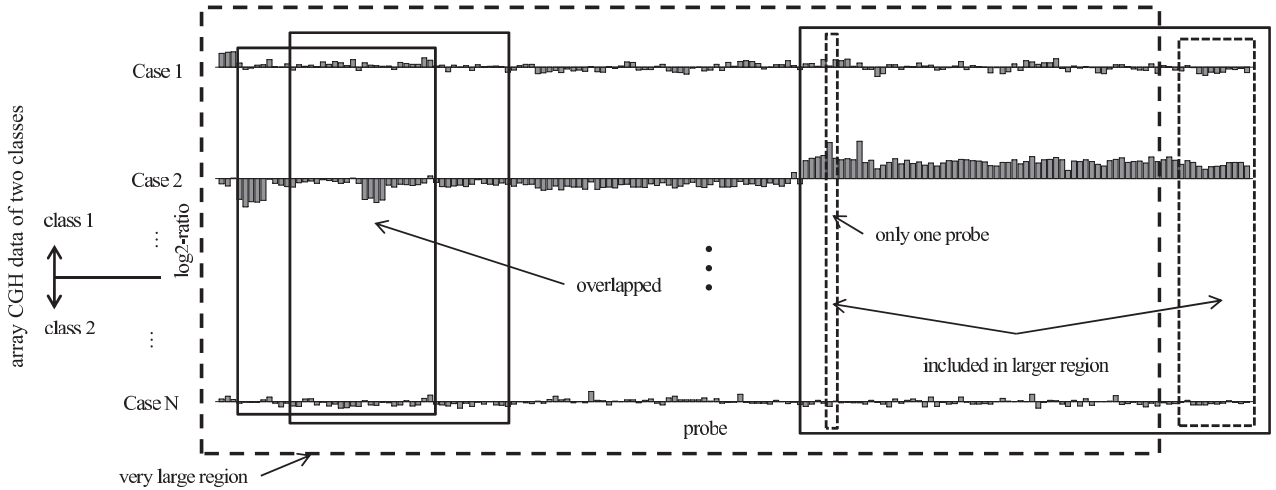


Fig. 2 Schematic illustration of candidates in aberrant region detection in array CGH data: Gray bars represent $\log_2$-ratios of each probe and the all rectangles denote candidate regions. There are many candidate regions with various lengths from regions consisting of a single probe to a region corresponding to the entire chromosome. Note that many regions are overlapped each other. We aim to identify statistically significant and differentially aberrant regions among these candidate regions.

2.2 Multivariate Test

To ensure the statistical reliability (such as p-values) of detected aberrant regions, we need statistical significance tests. For each region $\mathcal{R}_m$, $m = 1, \cdots, M$, we want to compute the statistical significance of the difference in the aberration patterns between group C_1 and group C_2. This problem is reduced to an $|\mathcal{R}_m|$-dimensional multivariate test for two samples $\{x_{icj}\}_{i\in\{1,\cdots,N|y_i=1\},(c,j)\in\mathcal{R}_m}$ and $\{x_{icj}\}_{i\in\{1,\cdots,N|y_i=2\},(c,j)\in\mathcal{R}_m}$ with sizes N_1 and N_2, respectively. In this multivariate test, the null hypothesis is that all the N $|\mathcal{R}_m|$-dimensional vectors $\{x_{icj}\}_{i\in\{1,\cdots,N\},(c,j)\in\mathcal{R}_m}$ are independently and identically distributed from a common $|\mathcal{R}_m|$-dimensional multivariate distribution, while the alternative hypothesis is that the two samples $\{x_{icj}\}_{i\in\{1,\cdots,N|y_i=1\},(c,j)\in\mathcal{R}_m}$ and $\{x_{icj}\}_{i\in\{1,\cdots,N|y_i=2\},(c,j)\in\mathcal{R}_m}$ are independently drawn from two different $|\mathcal{R}_m|$-dimensional multivariate distributions.

In statistics literature, parametric and nonparametric multivariate tests have been studied. If we assume multivariate Normal distributions and location shift difference in alternative hypothesis, we can use Hotelling T^2 test [12]. Hotelling T^2 test is a multivariate extension of t-test and it has the largest power when underlying assumptions are completely satisfied. Many multivariate distributions (probably including ours) do not exactly follow multivariate Normal distributions. In such cases, we can use nonparametric multivariate test. Many nonparametric multivariate test statistics are defined based on distances between pair of data points. For example, Ref. 13) extended a class of univariate nonparametric tests to multivariate one by constructing minimal spanning tree based on pairwise distances. Nearest neighbor test [8),9),14)] is also constructed under a similar concept. We call these types of multivariate test as *multivariate approach*.

These multivariate tests are not frequently used in practical applications, and a simpler approach (what we call *univariate approach* in contrast to multivariate approach) is often adopted. For quantifying the difference between $|\mathcal{R}_m|$-dimensional two samples, one can, for example, compute average of $|\mathcal{R}_m|$ univariate two-sample statistics. If ordinal two-sample t-test is used in this situ-

ation, a simple multivariate statistic is defined as $T_{\mathcal{R}_m} = \frac{1}{|\mathcal{R}_m|}\sum_{(c,j)\in\mathcal{R}_m} t_{cj}$, where t_{cj} is the t-value of the two univariate samples $\{x_{icj}\}_{i\in\{1,\cdots,n|y_i=1\}}$ and $\{x_{icj}\}_{i\in\{1,\cdots,n|y_i=2\}}$. Many other forms of univariate approach are possible. For example, Ref. 15) suggested to use the maximum t-statistics instead of the average, i.e., in the above problem setup, the statistic of the region $\mathcal{R}_m$ is given by $T_{\mathcal{R}_m} = \max_{(c,j)\in\mathcal{R}_m} t_{cj}$.

The advantage of such univariate approach is its simplicity both in interpretation and computation. If we use univariate approach for our task, we can first compute the t-statistic of each probe, and followed by computing the multivariate statistic for $\mathcal{R}_m$, $m = 1, \cdots, M$, simply by averaging the t-statistics in $\mathcal{R}_m$. On the other hand, if the multivariate data has correlation among variables, univariate approach is less powerful than multivariate approach. The computational cost of multivariate approach is usually much larger than univariate approach. Since we have a huge number of candidate regions M and we need to repeatedly compute the set of statistics in label permutation operation (see next subsection), computational burden in multivariate approach is a major limitation for our problem. On the other hand, multivariate approach would have larger power than univariate approach if the multivariate data has correlation among variables.

2.3 Multiple Testing

In differentially aberrant region detection problem, we have to consider a multiple testing problem since the large number of statistical tests is performed simultaneously. In multiple testing correction, we need to take correlation structure among test statistics into consideration. If the test statistics are independent and the multiplicity (the number of tests) is not so large, we can use several multiple testing correction procedures such as Bonferroni correction [16]. On the other hand, if the test statistics have complicated correlation structure or the multiplicity is large these off-the-shelf correction procedures are known to be too conservative (less powerful). In our differentially aberrant region detection problem the multiplicity is huge (i.e., we have M tests) and many regions are highly correlated because they are largely overlapped.

A practical and useful correction approach for large-scale and highly correlated multiple testing problem is to use label-permutation. Label permutation is a general procedure for estimating null distributions. In our problem setup, we first shuffle the labels $\{y_i\}_{i=1}^n$ randomly, and compute the multivariate statistic, say, $T_{\mathcal{R}_m}$, for each candidate region $\mathcal{R}_m$, for $m = 1, \cdots, M$. This process is repeated many times (for instance, 1000 times) so that the null distribution is estimated with sufficient accuracy. An important advantage of label permutation is that the null distribution is estimated without collapsing the correlation structure among regions. On the other hand, label permutation procedure is computationally quite demanding. If we repeat B permutation procedure in our problem, we need to repeat the computation of a statistic MB times.

Large-scale and highly correlated multiple testing problem has been intensively studied in expression microarray data analysis [6),7)]. Multiple comparison free measures frequently used in microarray studies are family-wise error (FWE) rate and false discovery rate (FDR). FWE is the probability of finding at least one false positive (committing at least one type I error), while FDR is the proportion of the false positives (falsely rejected hypotheses) among all the positives (rejected hypotheses). Using label permutation, both of FWE and FDR can be computed without collapsing the correlation structure among test statistics.

2.4 Comparing Aberrant Regions with Different Lengths

In differentially aberrant region detection problem, we perform statistical test of genomic regions with various different lengths. For example, in extreme case, we need to compare on an equal footing the statistical significances of a genomic region corresponding to a single probe and that consisting of the entire chromosome. Therefore, we need to use a multivariate statistic that is comparable among different sizes of regions. In other words, we need to introduce a multivariate statistic that does not depend on the dimensionality $|\mathcal{R}_m|$. Note that many multivariate statistics are dependent on the dimensionality. For example, the null distribution of the average t-statistic $T_{\mathcal{R}_m} = \frac{1}{|\mathcal{R}_m|}\sum_{(c,j)\in\mathcal{R}_m} t_{cj}$ would be different for different dimension $|\mathcal{R}_m|$, i.e., the variance of the null distribution would be smaller for larger regions. Note that a normalization or a standardization of the different scale of statistics from different dimensional data is possible only when each variable is independently and identically distributed or the correlation structure is completely known in advance.

3. Nearest-Neighbor Multivariate Test

In this section we introduce nearest-neighbor multivariate test for detecting differentially aberrant regions. As we described in the previous section, the statistical test for this problem should satisfy the following requirements:

(1) The test should be able to incorporate correlation among probes, which means, the test should be able to examine multiple consecutive probes simultaneously, i.e., it should be multivariate approach.

(2) The test statistic should not depend on the dimensionality of the length of genome regions $|\mathcal{R}_m|$.

(3) The computational cost for computing the statistic should be moderate in label permutation.

In the next subsection we will show that these requirements are satisfied with the use of nearest-neighbor multivariate test.

3.1 Nearest-Neighbor Multivariate Test

For each region $\mathcal{R}_m$, $m = 1, \cdots, M$, we compute the leave-one-out cross-validation (LOOCV) error of k-nearest neighbor (k-NN) classifier using $\log_2$-ratios $\{x_{icj}\}_{i\in\{1,\cdots,n\},(c,j)\in\mathcal{R}_m}$. Roughly speaking, small LOOCV error indicates that the region $\mathcal{R}_m$ is differentially aberrant between C_1 and C_2, while large LOOCV error suggests that the aberrant pattern in $\mathcal{R}_m$ is NOT so different between the two groups. If we just want the *ranking* of differentially aberrant regions, we can simply rank the M regions in the increasing order of LOOCV error. However, if we want the statistical significances (p-value, FWE, FDR, etc.) as well, we need to compute the null distributions of those LOOCV errors. As discussed in the previous section, we use label permutation for estimating the null distributions. Since we need to compute the k-NN LOOCV error many times, if we naively implement it, the computational cost of the entire process would be overwhelmingly large. We use a simple trick to efficiently compute the label permuted k-NN LOOCV error in the following algorithm:

Algorithm:

Differentially aberrant region detection by k-NN

Input: $\log_2$-ratios $\{x_{icj}\}_{i=1,\ldots,nc=1,\ldots,22,j=1,\ldots,l_c}$, labels $\{y_i\}_{i=1,\ldots,n}$, the number of label-permutations B, the number of neighbors k, a distance function d, a threshold for FDR (or FWE) θ, and the maximum number of aberrant regions γ.

1-1: For each chromosome $c = 1, \ldots, 22$, enumerate all possible regions $\mathcal{R}_m, m = 1, \ldots, M$. For each region $\mathcal{R}_m, m = 1, \ldots, M$, create n-by-k nearest-neighbor table T_m, and record the indices of the k "nearest" cases in the i-th row of the table T_m, where "nearest" is measured by the distance function d.

1-2: Compute the k-NN LOOCV error s^*_m for each region $\mathcal{R}_m, m = 1, \ldots, M$, based on the labels $\{y_i\}_{i=1}^n$. Set $b = 1$.

2: Permute the labels $\{y_i\}_{i=1}^n$ at random and let those labels be $\{y_i^{(b)}\}_{i=1}^n$. Compute the k-NN LOOCV error $s_m^{(b)}$ for each region $\mathcal{R}_m, m = 1, \ldots, M$, based on the labels $\{y_i^{(b)}\}_{i=1}^n$. If $b < B$ then $b \leftarrow b + 1$ and go back to 1-2.

3: For each region $\mathcal{R}_m, m = 1, \ldots, M$, compute

$$\mathrm{FDR}_m = \frac{\sum_{b=1}^B \sum_{m=1}^M I\{s^*_m \le s^{(b)}_m\}}{B \sum_{m'=1}^M I\{s^*_m \le s^*_{m'}\}},$$

$$\mathrm{FWE}_m = \frac{\sum_{b=1}^B \sum_{m=1}^M I\{s^*_m \le s^{(b)}_m\}}{B},$$

where $I\{\cdot\}$ is the indicator function.

4: Select the regions whose FDRs (or FWEs) are less than the threshold θ. If two or more overlapped regions are selected, select only the smallest one. If more than γ regions are selected, keep only the top γ regions.

Output: (at most γ) differentially aberrant regions with FDRs (or FWEs) less than θ.

In k-NN classification, we need to select the number of neighbors k and it often largely affects its classification performances. On the other hand, when we use k-NN for statistical testing, we can use the average of k-NN LOOCV error for several ks, such as $k = 1, 3$ and 5.

3.2 Advantages of the Algorithm

Nearest-neighbor method is a simple classification algorithm, but it often works well in real world applications. For example, in tumor classification problem using expression microarray data, it is reported that nearest-neighbor classifier showed the best performance among many classification algorithms including

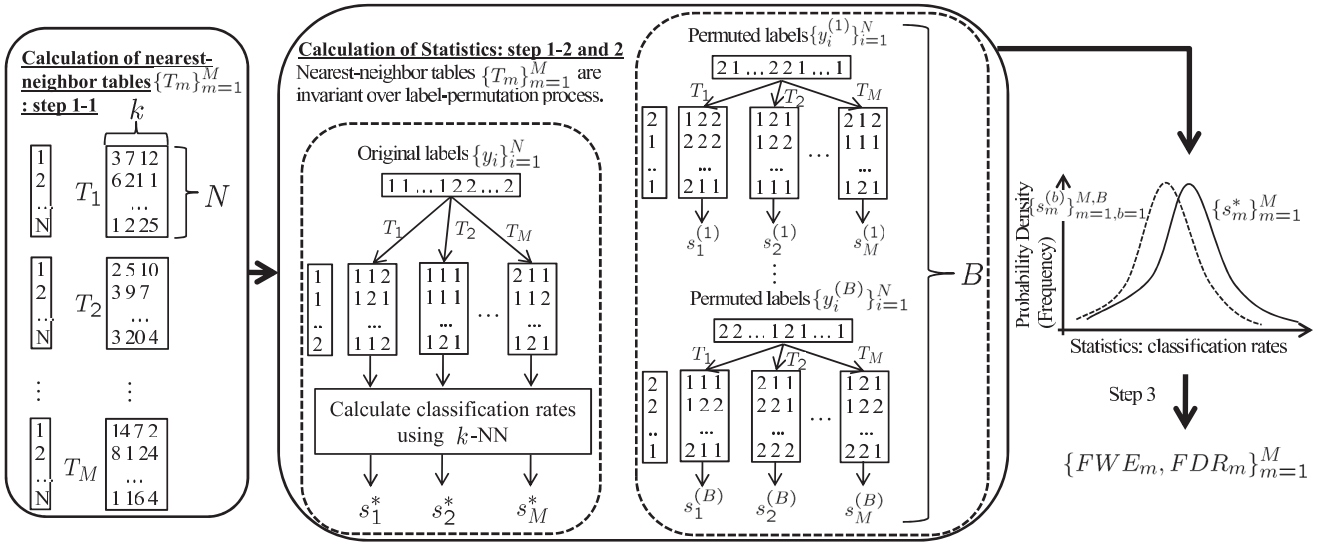


Fig. 3 Nearest-Neighbor multivariate test: In Step 1-1, we compute the nearest-neighbor tables. Noting that the nearest-neighbor tables are invariant under label permutations, we can reduce the computational cost for the label-permutation process. In Step 3 we estimate FDR and FWE for each candidate region based on the estimated null distributions.

decision tree, support vector machine, etc.[17]. It suggests that k-NN LOOCV error would be effective for measuring the classifiability of each candidate region, and it further means that our nearest neighbor test has large power to detect the differentially aberrant regions (i.e., small type II error). Therefore, we can say that our approach satisfies the first requirement 1) described in the previous section.

In addition, the above approach also satisfies the second requirement 2) because k-NN LOOCV error does not depend on the length of the region $|\mathcal{R}_m|$. On the other hand, a disadvantage of the LOOCV error is that it takes only a limited number of discrete values. It means that the LOOCV error is sometimes too granular to finely discriminate subtle differences, and many regions tend to have identical LOOCV error score. In this case, we need to employ some heuristics such as put priority on smaller regions.

Finally we emphasize the advantage of our approach for the third requirement 3). In general, the computational costs of multivariate approaches are larger than the costs of univariate approaches, and thus the multivariate approach is not suitable for the problem which has many candidates and requires many re-sampling iterations. However, our method can alleviate the computational cost using a simple trick. The computational cost of step 1 is rather larger because we need to compute the k nearest-neighbors of n arrays for all the M regions. Therefore, the entire cost would be huge if we naively repeat the step 1 for each label permutation. To alleviate this problem, we divide the step 1 into two steps: step 1-1 and 1-2. Note that, in the step 1-2, we can compute the k-NN LOOCV error s_m^* in the order $\mathcal{O}(kn)$ if we use the nearest-neighbor table $\{T_m\}_{m=1}^M$ which is already computed in the step 1-1. Although some readers might think that we have to repeat the steps 1-1 and 1-2 for each permuted labels $\{y_i^{(b)}\}$, we do not have to repeat the step 1-1 in each label permutation because we do not use label information when computing the nearest-neighbor table in the step 1-1.

In other words, the nearest-neighbor table is invariant under label permutation. Exploiting this fact, in each permutation, we only perform the step 1-2, and thus the order of computational cost for each region amounts only to $\mathcal{O}(kn)$. In our algorithm we need integer-type Mnk memory-space instead. This could be improved both in memory size and computation efficiency if we use bit operation in our implementation.

Figure 3 schematically illustrates our nearest neighbor multivariate test approach.

4. Application to Array CGH Data Analysis

In this section we apply the differentially aberrant region detection algorithm to previously published array CGH data set. Then we use the detected regions for tumor subtype classification task.

4.1 Data Set and Prepossessing

In this experiment we apply our algorithm to 75 BAC CGH arrays taken from 75 lymphoma patients. The patient sample were collected and investigated in Aichi Cancer Center [2),11),19)]. Among 75 cases, 46 cases were diagnosed (by a pathologist) as diffuse large B-cell lymphoma (DLBCL) and 29 cases were diagnosed as mantle cell lymphoma (MCL). The gene expressions of the 46 DLBCL cases were profiled and 18 cases were estimated as activated B-cell type (ABC) DLBCL and 28 cases were estimated as germinal center B-cell type (GCB) DLBCL [11)]. In our experiments we conducted two studies. In the first study, the goal is to detect differentially aberrant regions between 46 DLBCL cases and 29 MCL cases. In the second study, we aim to identify differentially aberrant regions between 28 ABC subtype cases and 18 GCB subtype cases. Each array has 2035 BAC probes. Figure 1 shows some examples of $\log_2$-ratio sequences from these arrays.

The $\log_2$-ratio signals in each array were standardized such that the median takes 0. This standardization yields a small bias, i.e., arrays having many amplifications are biased downward and arrays having many deletions are biased upward. To identify gain and loss aberrant regions individually, we analyzed *gain* $\log_2$*-ratio sequence* and *loss* $\log_2$*-ratio sequence* separately. To setup gain $\log_2$-ratio sequences, $\log_2$-ratios less than 0.1 values are replaced by a random value in $[0.0, 0.1]$ and loss $\log_2$-ratio sequences are made by replacing $\log_2$-rations greater than -0.1 by a random value in $[-0.1, 0.0]$.

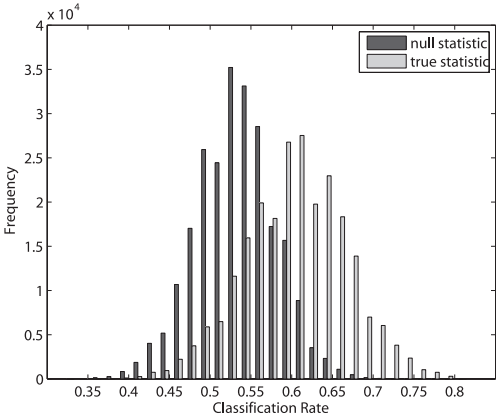


Fig. 4 Histograms of the statistics and the null distribution: The peak of the null distribution is around 0.5, i.e., the LOOCV *classification accuracy* of permuted data set is around 0.5, while the statistics, LOOCV *classification accuracy* of the original (non-permuted) data set are biased upward.

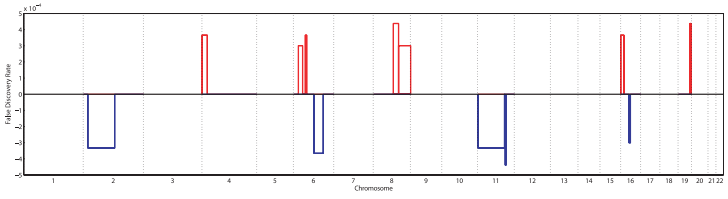
4.2 Differentially Aberrant Region Detection

The FDR threshold θ was set to be 0.0005 for the first DLBCL vs. MCL study, and 0.005 for the second ABC vs. GCB study.

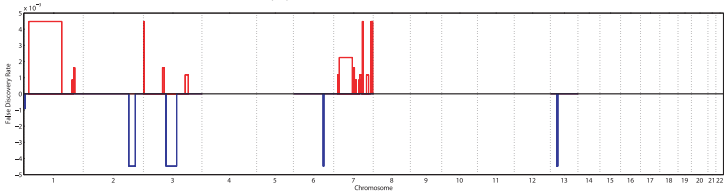
The number of label permutations B was set to be 1000. First we generated the nearest-neighbor tables $\{T_m\}_{m=1}^{M}$ in Step 1-1, and then calculated the set of test statistics $\{s_m^*\}_{m=1}^{M}$ and their null versions $\{s_m^{(b)}\}_{m=1,b=1}^{M,B}$ in Step 1-2 and Step 2, respectively. **Figure 4** shows the histograms of the set of statistics $\{s_m^*\}_{m=1}^{M}$ and their null distributions $\{s_m^{(b)}\}_{m=1,b=1}^{M,B}$. Note that, in this experiment, we measured these values by LOOCV *classification accuracy* instead of LOOCV error, which means, the larger values are better.

Using the estimated null distributions we estimated the FDR (or FWE) for each of M regions, and those having FDRs (or FWEs) less than the threshold θ were detected as differentially aberrant regions. **Figure 5** shows the detected regions for (a) DLBCL vs. MCL study and for (b) ABC vs. GCB study.

These results indicate that our proposed method could detect both of small



(a)DLBCL vs. MCL



(b)ABC vs. GCB

Fig. 5 Detected aberrant regions: The Horizontal axis indicates chromosome and the vertical axis indicates the value of FDR. Red lines show gain aberrant regions, while blue lines show loss aberrant regions. Note that, FDRs of loss aberrant regions are indicated in negative values, that is, −FDR.

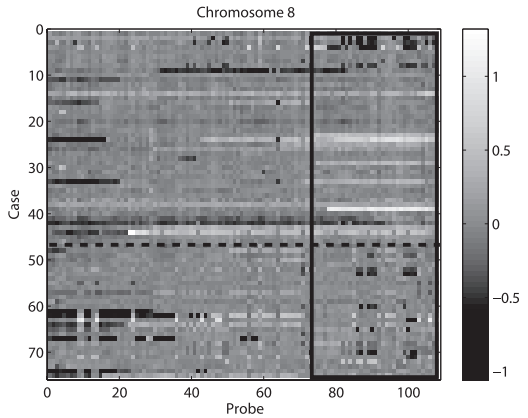


Fig. 6 Example of detected aberrant region (Chromosome 8): The smallest FDR region of DLBCL / MCL array CGH data is indicated by a solid-line rectangle. The vertical axis denotes BAC probes, while the horizontal axis denotes cases. Cases in the upper side of the horizontal dot line are DLBCL patients while in the lower side of the line are MCL patients. Cells with the greater $\log_2$-ratio values are shown in white color and ones with the smaller values are shown in black.

and large regions. **Figure 6** shows $\log_2$-ratios in a detected aberrant region. In the detected aberrant region (surrounded by solid lines in the figure), we can see that the DLBCL cases show more gains (white) than the MCL cases. These results demonstrate that our approach is a potential to find important genomic regions for differentiating tumors.

4.3 Tumor Classification Using the Detected Regions

In this subsection we use the detected regions for tumor classification task. To evaluate the classification performance, we used LOOCV. First we leave one array out and detect aberrant regions using the rest of the arrays. Then we estimate the posterior probability of the left-out array using the *voting* by the detected regions. Each detected region has one vote, and the vote is determined by k-nearest neighbor classification with the distance computed using the $\log_2$-ratios in that region. For example, if 10 regions are detected as differentially aberrant, and three regions vote to DLBCL and seven regions vote to MCL, then the DLBCL probability is said to be 0.3 and the MCL probability is said to be 0.7.

We compared the classification performance with ADM [18] and CLAC [3]. ADM was a standard method to analyze array CGH data, however, it was proposed to identify aberrant regions for one array. Therefore, we first compute t-value of each probe from the two samples and they are used as the input to ADM. CLAC was also originally proposed to identify aberrant regions for one array (see Section 1). Thus we use t-value calculated in the same way as ADM case as the input to CLAC. Note that these modified ADM and CLAC are instances of univariate approach (see Section 2). In this tumor classification task, we set the FDR threshold θ to be 0.00125 for DLBCL vs. MCL study and 0.0125 for ABC vs. GCB study. Since ADM does not provide FDR, a significant level of ADM is defined as the p-value with Bonferroni correction. That is, in our case, the total number of candidate regions is 118328, therefore, the thresholds for ADM are 0.00125/118328 for DLBCL vs. MCL study and 0.0125/118328 for ABC vs. GCB study respectively.

We measure the classification accuracy by the ROC curve and the AUC. The ROC (receiver operating characteristic) curve represents the sensitivity vs. specificity for a binary classifier as its threshold being varied. In DLBCL vs. MCL

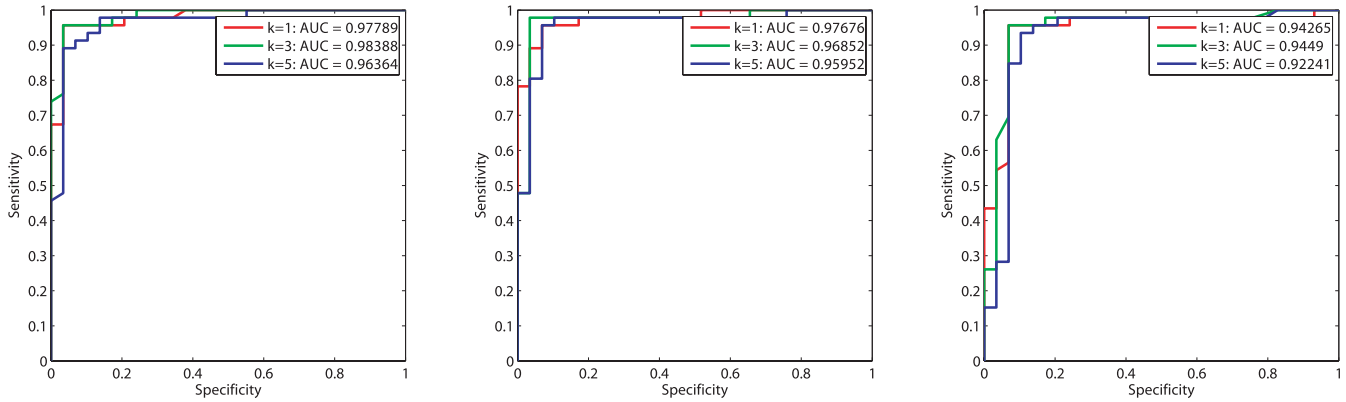


Fig. 7 ROC curve for DLBCL / MCL classification ($k = 1, 3, 5$): nearest neighbor multivariate (left), ADM (center) and CLAC (right).

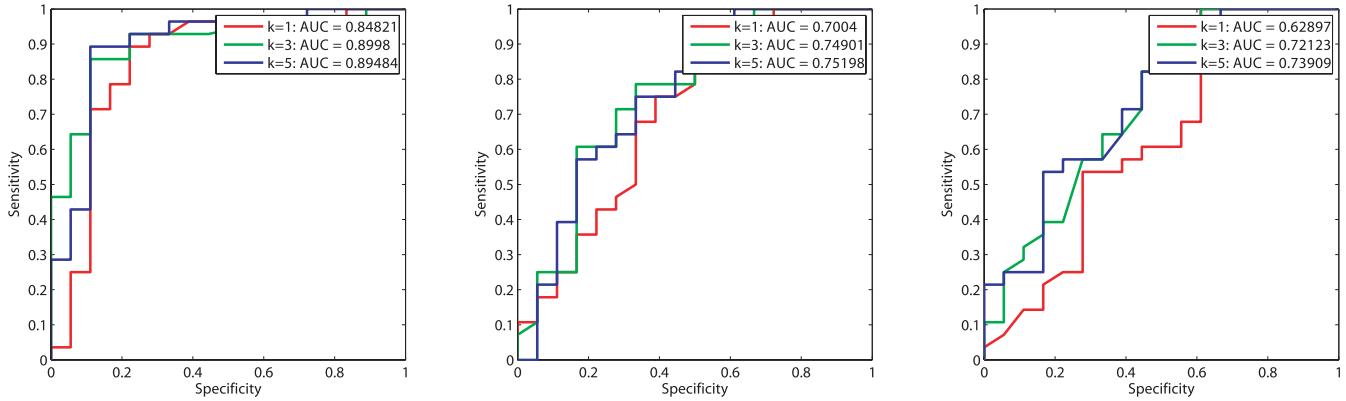


Fig. 8 ROC curve for ABC / GCB classification ($k = 1, 3, 5$): nearest neighbor multivariate (left), ADM (center) and CLAC (right).

study, we considered DLBCL is positive and MCL is negative, while in ABC vs. GCB study, ABC and GCB are considered as positive and negative, respectively. In **Figs. 7** and **8**, the horizontal axis denotes the false positive rate, while the vertical axis denotes the true positive rate. The AUC (area under curve) is defined as the area under the ROC curve, and it is used to evaluate the classification performance in a variety of cost (the relative costs between false positives and false negatives) as a single quantity. In ROC plane, the point $(0, 1)$ indicates the perfect classifier (no false positives and no false negatives), and AUC has the maximum value 1 in that case. The results are shown in Fig. 7 and Fig. 8. We evaluate with $k = 1, 3, 5$ in the classification process described above.

In DLBCL vs. MCL study, all the methods (ours, ADM and CLAC) show excellent performances. On the other hand, in ABC vs. GCB study, the performances of all the methods get worse compared with the first DLBCL vs. MCL study. This result is related to the fact that the distinction between ABC and

GCB subtypes is still vague in medical/biological literature[11]. Comparatively speaking, our method shows rather better performance than ADM and CLAC, and the difference of the performance is more remarkable in ABC vs. GCB study. From these results, our nearest neighbor multivariate approach has a potential to find regions that differentiate the two or more diseases or subtypes.

5. Conclusion

In this paper we study the problem of detecting differentially aberrant regions from two or multiple samples using nearest neighbor multivariate test. Our algorithm has several advantages. First, the algorithm can deal with various sizes of aberrant regions (from a single probe region to the entire genome region) in a unified framework. Second, we can compute the multiple comparison free significance measure such as FDR and FWE in a relatively small computational cost. Finally, our algorithm is multivariate approach (see Section 2), and thus it has a potential to incorporate spatial correlation among probes when detecting the differentially aberrant regions. An application to 75 lymphoma CGH arrays demonstrated the effectiveness of our approach.

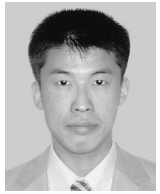
References

1) Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D. and Brown, P.O.: Genome-wide analysis of DNA copy-number changes using cDNA microarrays, *Nature Genetics*, Vol.23, No.1, pp.41–46 (1999).
2) Tagawa, H., Tsuzuki, S., Suzuki, R., Karnan, S., Ota, A., Kameoka, Y., Suguro, M., Matsuo, K., Yamaguchi, M., Okamoto, M., Morishima, Y., Nakamura, S. and Seto M.: Genome-wide array-based comparative genomic hybridization of diffuse large-b-cell lymphoma: Comparison between cd5-positive and cd5-negative cases, *Cancer Research*, Vol.64, pp.5948–5955 (2004).
3) Wang, P., Kim, Y., Pollack, J., Narasimhan, B. and Tibshirani, R.: A method for calling gains and losses in array CGH data, *Biostatistics*, Vol.6, pp.45–58 (2005).
4) Olshen, A.B. and Venkatraman, E.S.: Circular binary segmentation for the analysis of array-based DNA copy number data, *Biostatistics*, Vol.5, No.4, pp.557–572 (2004).
5) Lipson, D., Aumann, Y., Ben-dor, A., Linial, N. and Yakhini, Z.: Efficient calculation of interval scores for dna copy number data analysis, *Journal of Computational Biology*, Vol.13, No.2, pp.215–228 (2006).
6) Kong, S.W., Pu, W.T. and Park, P.J.: A multivariate approach for integrating genome-wide expression data and biological Knowledge, *Bioinformatics*, Vol.22, No.19, pp.2373–2380 (2006).
7) Ge, Y., Dudoit, S. and Speed, T.P.: Resampling-based multiple testing for microarray data analysis, *Statistician*, Vol.45, No.4, pp.407–436 (1996).
8) Henze, B.: A multivariate two-sample test based on the number of nearest neighbor type coincidences, *Annals of Statistics*, Vol.16, No.2, pp.772–783 (1988).
9) Schilling, M.H.: Multivariate two-sample tests based on nearest neighbors, *J. Am. Stat. Assoc.*, Vol.81, pp.799–806 (1986).
10) Ota, A., Tagawa, H., Karnan, S., Tsuzuki, S., Karpas, A., Kira, S., Yoshida, Y. and Seto, M.: Identification and characterization of a novel gene, c13orf25, as a target for 13q31-q32 amplification in malignant lymphoma, *Cancer Research*, Vol.64, pp.3087–3095 (2004).
11) Tagawa, H., Karnan, S., Suzuki, R., Matsuo, K., Zhang, X., Ota, A., Morishima, Y., Nakamura, S. and Seto, M.: Genome-wide array-based cgh for mantle cell lymphoma: Identification of homozygous deletions of the proapoptotic gene bim, *Oncogene*, Vol.24, pp.1348–1358 (2005).
12) Hotelling, H.: The generalization of student's ratio, *Annals of Mathematical Statistics*, Vol.2, pp.360–378 (1931).
13) Friedman, J.H. and Rafsky L.C.: Multivariate generalizations of the Wald-Wolfovits and Smirnov two-sample test, *Annals of Statistics*, Vol.7, pp.697–717 (1979).
14) Hall, P. and Tajvidi, N.: Permutation tests for equality of distributions in high-dimensional settings, *Annals of Statistics*, Vol.27, pp.1385–1414 (1999).
15) Boyett, J.M. and Shuster, J.J.: Nonparametric one-sided tests in multivariate analysis with medical applications, *Journal of the American Statistical Society*, Vol.72, pp.665–668 (1977).
16) Holm, S.: A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics*, Vol.6, pp.65–70 (1979).
17) Dudoit, S., Fridlyand, J. and Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data, *J. Am. Stat. Assoc.*, Vol.97, No.457, pp.77–87 (2002).
18) Lipson, D., Aumann, Y., Ben-Dor, A., Linial, N., and Yakhini, Z.: Efficient calculation of interval scores for DNA copy number data analysis, *Journal of Computational Biology*, Vol.13, No.2, pp.215–228 (2006).
19) Takeuchi, I., Tagawa, H., Tsujikawa, A., Nakagawa, M., Suguro, M.K., Guo, Y. and Seto, M.: The potential of copy number gains and losses, detected by array-based comparative genomic hybridization, for computational differential diagnosis of b-cell lymphomas and genetic regions involved in lymphomagenesis, *Haematologica*, Vol.94, pp.51–69 (2009).

(Received May 6, 2010)
(Accepted July 9, 2010)
(Released October 13, 2010)

(Communicated by *Wataru Fujibuchi*)

Yuta Ishikawa received his B.E. and M.E. degrees in computer science from Nagoya Institute of Technology in 2006 and 2008, respectively. He is currently pursuing the Ph.D. degree from the Department of Engineering, Nagoya Institute of Technology. His current research interests are statisitcal pattern recognition and its application to bioinformatics. He is a student member of the Institute of Electronics, Information and Communication Engineers and Japan Statistical Society.



Ichiro Takeuchi is an associate professor at Nagoya Institute of Technology, Japan. He received B.E., M.E., and D.E. degrees from Nagoya University, Japan, in 1996, 1998, and 2000, respectively. His research interests include machine learning, optimization, and bioinformatics.