BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 18 2010, pages 2313–2320
doi:10.1093/bioinformatics/btq428

Systems biology Advance Access publication July 21, 2010

Integrative classification and analysis of multiple arrayCGH datasets with probe alignment

Ze Tian and Rui Kuang*
Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, MN, USA
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Array comparative genomic hybridization (arrayCGH) is widely used to measure DNA copy numbers in cancer research. ArrayCGH data report log-ratio intensities of thousands of probes sampled along the chromosomes. Typically, the choices of the locations and the lengths of the probes vary in different experiments. This discrepancy in choosing probes poses a challenge in integrated classification or analysis across multiple arrayCGH datasets. We propose an alignment-based framework to integrate arrayCGH samples generated from different probe sets. The alignment framework seeks an optimal alignment between the probe series of one arrayCGH sample and the probe series of another sample, intended to find the maximum possible overlap of DNA copy number variations between the two measured chromosomes. An alignment kernel is introduced for integrative patient sample classification and a multiple alignment algorithm is also introduced for identifying common regions with copy number aberrations.

Results: The probe alignment kernel and the MPA algorithm were experimented to integrate three bladder cancer datasets as well as artificial datasets. In the experiments, by integrating arrayCGH samples from multiple datasets, the probe alignment kernel used with support vector machines significantly improved patient sample classification accuracy over other baseline kernels. The experiments also demonstrated that the multiple probe alignment (MPA) algorithm can find common DNA aberrations that cannot be identified with the standard interpolation method. Furthermore, the MPA algorithm also identified many known bladder cancer DNA aberrations containing four known bladder cancer genes, three of which cannot be detected by interpolation.

Availability: http://www.cs.umn.edu/compbio/ProbeAlign

Contact: kuang@cs.umn.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 26, 2010; revised on June 30, 2010; accepted on July 18, 2010

1 INTRODUCTION

It has been confirmed in many recent studies that aberrations in chromosome copy number, rearrangement and structures have association with the disease (Feuk *et al.*, 2006). Among these chromosomal aberrations, DNA copy number variations (CNVs), the events of amplification or deletion of a large DNA segment on chromosomes, are believed to play an important role in tumorigenesis (Redon *et al.*, 2006; Shlien and Malkin, 2009). Chromosome CNVs can be measured by comparative genomic hybridization (CGH), which compares the copy number of a differentially labeled case sample with a reference DNA from a normal individual. ArrayCGH technology based on DNA microarray can currently allow genome-wide identification of regions with CNVs at different resolutions (Carter, 2007). The arrayCGH data was used to discriminate healthy patients from cancer patients and classify patients of different cancer subtypes. Thus, arrayCGH data is considered as a new source of biomarkers that provide important information of candidate cancer loci for the classification of patients and discovery of molecular mechanisms of cancers (Sykes *et al.*, 2009).

ArrayCGH allows rapid mapping of DNA CNVs of a tumor sample by a locus-by-locus measure. In an arrayCGH experiment, probes (short DNA segments on chromosomes) of different lengths and locations are chosen for comparative genomic hybridization. The CNV information at a probe location is reported by the log-ratio of the probe intensity between a case and a reference. The sizes and the number of the probe determines the resolution of an arrayCGH experiment. The length of the DNA probes used by current platforms ranges from 100 bp to 5 kb, and in a typical study, the number of probes ranges from several hundreds to tens of thousands. For example, a Human array 2.0 chromium surface array can consist of only 2464 probes at 1.5 Mb resolution, whereas a high-resolution tiling array provides measurements of 36 288 probes.

The large difference in the sizes and numbers of probes makes integration of multiple arrayCGH datasets used for similar studies, a challenging problem. From a data analysis perspective, the arrayCGH data are series of real values labeled by their chromosomal positions. The traditional approach is an interpolation between the two series of probe locations from two platforms. As described in Figure 1, in an interpolation probe locations on one platform are added (interpolated) to the target platform and the corresponding CNV intensities are guessed by intensities of the most nearby probes in the target platform. However, when the two platforms have large variations in the number of probes, the interpolation is not an appropriate way to integrate the two series of probes. Specifically, when the two series are both sparse, the interpolated points might give misleading or wrong information. For example, in Figure 1, some of the probes are interpolated to locations close to probes with opposite CNV. When the probes are sparsely located, this might introduce false positives. Moreover, some of the interpolated probes are far from other probes. In these cases (i.e. the interpolated probes marked by '?' in the figure), it is ambiguous to decide the CNV for the interpolated new probes.

*To whom correspondence should be addressed.

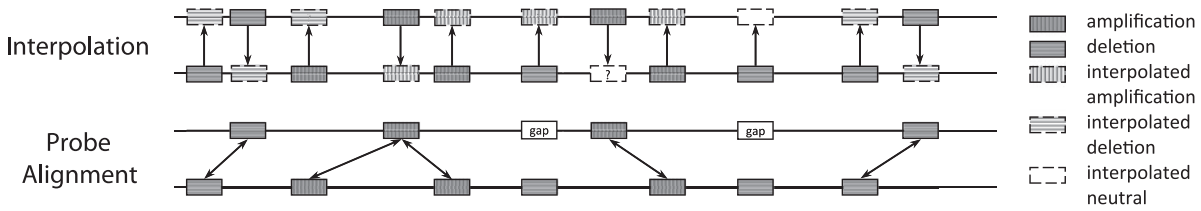

Fig. 1. A hypothetical example of interpolation and probe alignment. Interpolation and probe alignment are applied to compare the same two chromosomes. The probe locations are marked with vertically striped blocks (amplifications) or horizontally striped blocks (deletions). The interpolation interpolates each probe to its corresponding position on the chromosome and add a new probe in the other series. The CNV value of the new probe is guessed by checking the neighboring probes in the same series. The probe alignment aligns nearby probe pairs in the two series with similar CNVs. The matched probe pairs are connected by arrows in the alignment. The interpolated probes marked with '?' are possibly wrongly labeled in the interpolation.

In this article, we propose to integrate arrayCGH datasets with probe alignment. As shown in Figure 1, two series of probes from two arrayCGH samples are aligned based on both their chromosomal locations and their CNV log-ratios. The probes in one series are matched to the probes in the other. The alignment representing an approximate positional matching of the two compared chromosomes with the maximum possible overlap of CNV events. The probe alignment seeking the maximum possible CNV overlap is motivated by the fact that two tumor samples are likely to share some common aberration regions related to tumor development and progression. If a probe in one series shares the same CNV with another nearby probe in the other series, it is likely that the two probes capture the same common tumor aberration in the two samples. Thus, the two probes should be aligned to enhance the signal. Based on this assumption, a probe alignment is more capable of detecting weak common CNV signals between two chromosomes if only sparse information is available. The problem of finding the best alignment of two probe series can be solved by a variation of the standard global sequence alignment algorithm (Durbin *et al.*, 1998). Compared with interpolation, the main advantage of probe alignment is that the copy number of the chromosomal regions between probes are inferred based on the information of all the probes, instead of only the nearby ones, and the optimal global alignment provides the best guess of the CNVs in these regions.

Based on probe alignment, a probe alignment kernel derived from the probe alignment scores is introduced for classification of patient samples from multiple datasets. In the classification task, a binary classifier is trained to predict the tumor grade or cancer stages of a patient using the CNV information in different arrayCGH data. With the probe alignment kernel, many more patient samples from other datasets can be used to improve classification performance on one dataset. A multiple alignment algorithm is also designed to align all the probe series for identifying common CNV regions across patient samples from many datasets. In this case, many probe series from several arrayCGH datasets are aligned probe-by-probe, and the final alignment profile can reveal the detected common CNV regions along the chromosomes.

Sequence alignment algorithms are well-established methods for protein/nucleic acid sequence analysis (Durbin *et al.*, 1998). Famous examples are Needleman–Wunsch algorithm, Smith–Waterman algorithm and fast approximations such as BLAST/PSI-BLAST (Altschul *et al.*, 1997). The applications of these algorithms were relatively only limited to biological sequences rather than other types of data. It is worth mentioning that Aach and Church (2001) proposed to use an alignment algorithm to align time series of gene expressions. Although motivated by another application, the central philosophy of finding a synchronization of two series to replace simple interpolation is closely related. There are several previous approaches for segmentation and CNV detection from arrayCGH data. For example, Guha *et al.* (2008) proposed a Bayesian hidden Markov model to model relation between neighboring probes. Because these approaches assumes discrete copy number states without quantifying the distance between probes, they are not directly applicable in the multi-platform scenario.

2 METHODS

In this section, we first describe how to align two probe series. We next introduce the probe alignment kernels for classification of CNV samples and a multiple alignment algorithm for identifying common disease in CNV regions. Finally, the time complexity of probe alignment is analyzed.

2.1 Pairwise alignment of probe series

We denote the series of arrayCGH probes on a chromosome as a finite sequence of tuples $(x_1,l_1),(x_2,l_2),\ldots,(x_i,l_i),\ldots$, where each x_i denotes the log-ratio intensity of the i-th probe and l_i denotes the location of the probe on the chromosome by kilo base pairs (kb). Given two such sequences $U=(u_1,a_1),(u_2,a_2),\ldots,(u_n,a_n)$ and $V=(v_1,b_1),(v_2,b_2),\ldots,(v_m,b_m)$ of length n and m, respectively, we define three functions, $S(i,j)$, $L(i,j)$ and $R(i,j)$, for computing the alignment between the subsequences $(u_1,a_1),\ldots,(u_i,a_i)$ in U and the subsequences $(v_1,b_1),\ldots,(v_j,b_j)$ in V. $S(i,j)$ is the optimal alignment score when (u_i,a_i) is aligned with (v_j,b_j); $L(i,j)$ is the optimal alignment score when (u_i,a_i) is aligned with a gap and $R(i,j)$ is the optimal score when (v_j,b_j) is aligned with a gap. Finally, a function $M(i,j)$ defined as the max of $S(i,j)$, $L(i,j)$ and $R(i,j)$ gives the optimal alignment score up to position i in U and position j in V. With initialization $S(0,*)=S(*,0)=L(*,0)=L(0,*)=R(*,0)=R(0,*)=0$, the functions can be recursively evaluated with a variation of standard dynamic programming as follows,

$$S(i,j) = \max\begin{cases} M(i-1,j-1)+s(i,j) \\ S(i,j-1)+s(i,j) \\ S(i-1,j)+s(i,j) \end{cases}$$

$$L(i,j) = M(i-1,j)+g$$

$$R(i,j) = M(i,j-1)+g$$

$$M(i,j) = \max\{S(i,j),L(i,j),R(i,j)\}$$

where function $s(i,j)$ gives the substitution score between (u_i,a_i) and (v_j,b_j) and constant g is a gap penalty. The gaps in the alignment are introduced to capture the regions that are sufficiently covered by probes in one platform but not the other. Note in our formulation, a probe in one probe series can be matched with multiple probes in the other series, i.e. we also consider $S(i,j-1)+s(i,j)$ and $S(i-1,j)+s(i,j)$ when $S(i,j)$ is calculated. Because

different platforms have different probe densities, it is more reasonable to allow one-to-many matching in the alignment.

Given two probe series U and V, the substitution score $s(i,j)$ between (u_i,a_i) and (v_j,b_j) needs to be designed to quantify a composition of two measurements: first, how close the two positions a_i and b_j are, and second, how similar the two CNV log-ratios u_i and v_j are. The substitution scores should encourage alignment of probe pairs that have similar amplification/deletion log-ratios at nearby locations. A simplest scoring function can be defined as follows:

$$s(i,j)=e^{\frac{-|a_i-b_j|}{\sigma}}*u_i*v_j, \tag{1}$$

where $e^{\frac{-|a_i-b_j|}{\sigma}}$ quantifies how close the two positions are and u_i*v_j is a simple measure of whether the two CNV log-ratios are similar or opposite. In this scoring function, the distance on a chromosome is normalized by a constant σ, which can be estimated from the actual probe locations in the probe series data. The closer the two probes, the larger the score. The similarity between two CNV log-ratios is simply taken as the product of the two values. The main advantage of this scoring function is that it is straightforward and parameter free. Note that, because the probes might represent the DNA copy numbers in chromosome regions of various lengths, the optimal probe alignment does not necessarily preserve the best sequential mappings between the two series by locations. Thus, a probe might not be matched with its closest peers in the other series in the alignment.

To produce more refined multiple probe alignments (MPAs), the scoring function in Equation (1) can be extended to be a positive function as follows:

$$s(i,j)=e^{\frac{-|a_i-b_j|}{\sigma}}*([u_i*v_j]_+ +1), \tag{2}$$

where $[u_i*v_j]_+$ is a refined product based on the sign and the value of u_i and v_j, defined as below

$$[u_i*v_j]_+ = \begin{cases} u_i*v_j, & \text{if } u_i*v_j \geq 0. \\ 0, & \text{otherwise.} \end{cases}$$

The product is then shifted by 1 to keep the positiveness. The refined similarity between u_i and v_j considers two scenarios. First, when u_i and v_j have the same sign, the similarity is still the product between them. Second, when u_i and v_j have different signs, the similarity between them is 0. In this scoring scheme, two probes with exactly the same position but different CNV signs will still be matched together to prevent needlessly penalizing the matches with gap insertions. This also implies that two probes at the same chromosomal position will still be matched even if they have opposite CNV events.

The choice of the gap penalty g depends on the substitution function $s(i,j)$. If the substitution function in Equation (1) is used, g can simply be set to a very small positive constant, which guarantees that a gap is preferred over matching two probes with different signs, and that only a very small value is added to the overall alignment score from the gap insertions. Intuitively, when g is small enough, an identical alignment and similar kernels will be resulted. If the substitution function in Equation (2) is employed, the gap penalty $g=e^{\frac{-\tau}{\sigma}}$ is used to match the scaling of the new scoring function. By definition, this gap penalty only allows two probes with different signs to match if their distance on the chromosome is less than τ. This means that two probes with different signs will be matched if they are close-by sufficiently on the chromosome. τ can be empirically chosen as a value smaller than σ.

2.2 Probe alignment kernel

To address the problem that there is no direct way to integrate arrayCGH/CGH profiles from different platforms in classification, we propose a kernel based on probe alignment, which measures the similarity between samples with their probe alignment scores. The probes located on the DNA of each chromosome of human genome are aligned and the alignment scores are summarized as the kernel value. The alignment kernel function computing the best global alignment score between U and V is given as follows:

$$\begin{aligned} K(U,V) &= K(((u_1,a_1),\ldots(u_n,a_n)),((v_1,b_1),\ldots(v_m,b_m))) \\ &= \frac{M(m,n)}{m+n}, \end{aligned} \tag{3}$$

where the alignment score is normalized by $m+n$, the lengths of the two series to avoid the bias by the number of probes.

The kernel function $K(U,V)$ is used to compute the alignment score between two probe series of a particular human chromosome. Given two arrayCGH samples each with P chromosomes in genome (23 pairs of chromosomes in Human genome), the total alignment score between the two samples is the summation of the best global alignment scores of the P pairs of chromosomes. Let $\{U^{(1)},\ldots U^{(P)}\}$ and $\{V^{(1)},\ldots V^{(P)}\}$ denoting the two sets of probe sequences of the chromosomes. The probe-alignment kernel function $\mathcal{K}$ is defined as

$$\mathcal{K}(\{U^{(1)},\ldots U^{(P)}\},\{V^{(1)},\ldots V^{(P)}\})=\sum_{i=1}^{P}K(U^{(i)},V^{(i)})$$

We choose the scoring function given in Equation (1) for the probe alignment kernel. When all the probe positions can be perfectly matched between the two series, the probe alignment kernel tends to be very close to a simple linear kernel between the two probe series, where the kernel is simply the summation of the matched u_i*v_j in the two series. Note that the probe alignment kernel is not positive semi-definite (PSD), similar to the alignment kernel for protein classification (Liao and Noble, 2003). A positive constant is added to the diagonal of the kernel matrix to make it a valid kernel.

2.3 Multiple alignment of probe series

Similar to the multiple alignment algorithms for protein/DNA sequences, a multiple alignment procedure can also be used to align multiple probe series from different arrayCGH datasets. The MPA can reveal common amplification or deletion events even if the probe sets are at different chromosomal positions. We adopted the progressive multiple alignment strategy (Thompson *et al.*, 1994) and used the pairwise probe alignment with scoring function defined in Equation (2) for base alignment. The algorithm is described in Supplementary Section 1.

The algorithm continuously merges the two most similar probe series and finally, gives the alignment merged from all the series. Merging two probe series is the crucial step. To accomplish the best merging, we slightly extended the definition of a probe series by adding a weight to each probe in the series to make a new sequence of triples (x_i,l_i,w_i). Each weight w_i quantifies the importance of the probe by keeping the count of the number of probes that have been merged into this probe in previous alignments. Specifically, two such series $X_a=(u_1,a_1,w_1),\ldots,(u_n,a_n,w_n)$ and $X_b=(v_1,b_1,w'_1),\ldots,(v_m,b_m,w'_m)$ are merged as a new series $X_{a,b}$ based on their alignment. There are two types of probes in the alignment: (i) the probe is aligned with a gap and (ii) the probe is matched with some probe(s). New probes are calculated for these two cases as follows:

(1) If a triple (u_i,a_i,w_i) or (v_j,b_j,w'_j) is matched with a gap in the alignment between X_a and X_b, the triple is kept unchanged and directly added into the new series $X_{a,b}$.

(2) Because some consecutive probes in one series can be matched with some consecutive probes in the other series, multiple triples from both series might need to be merged as one new probe. Suppose $\{(u_i,a_i,w_i),(u_{i+1},a_{i+1},w_{i+1}),\ldots,(u_{i+k_1},a_{i+k_1},w_{i+k_1})\}$ and $\{(v_j,b_j,w'_j),(v_{j+1},b_{j+1},w'_{j+1}),\ldots,(v_{j+k_2},b_{j+k_2},w'_{j+k_2})\}$ are the two sets of multiple triples of length k_1 and k_2, respectively, that need to be merged. We combine them as a new tuple (z,c,w'') by

$$z=\sum_{p=i}^{i+k_1}u_p+\sum_{q=j}^{j+k_2}v_q;\ w''=\sum_{p=i}^{i+k_1}w_p+\sum_{q=j}^{j+k_2}w'_q;$$

$$c=\frac{\sum_{p=i}^{i+k_1}a_p*w_p+\sum_{q=j}^{j+k_2}b_q*w'_q}{w''}.$$

Note that the position of the new probe is calculated as a weighted average of the merged positions to relieve possible bias introduced by outlier probes.

2.4 Fast probe alignment in linear time complexity

The direct implementation of the recursive dynamic programming runs in quadratic time. Under an idea similar to fast-banded sequence alignment, probe alignment can be computed in linear time complexity. The motivation is that those probe pairs which are too far from each other will never be matched in the optimal alignment; otherwise, there always exists a better path that replaces the bad matches with more insertions and deletions. Thus, in the linear time implementation, only those probe pairs that are in locations close enough need to be considered in the alignment. Specifically, we prove the following two propositions:

Given two probe series, $U=(u_1,a_1),(u_2,a_2),\ldots,(u_n,a_n)$ and $V=(v_1,b_1),(v_2,b_2),\ldots,(v_m,b_m)$, and a small positive gap penalty g, let $u_{\max}=\max(|u_1|,\ldots,|u_n|)$ and $v_{\max}=\max(|v_1|,\ldots,|v_m|)$.

PROPOSITION 1. *The optimal alignment between U and V will only consist of gaps and pairs of aligned probes (U_i,V_j) with $|a_i-b_j|<\theta$, where $\theta=\sigma*(\ln(u_{\max}v_{\max})-\ln g)$.*

PROPOSITION 2. *The number of probe pairs that need to be considered in probe alignment is $O(\frac{\theta m}{\delta})$, where $\delta=\min\{b_{j+1}-b_j\}$.*

Proposition 1 states that in the optimal alignment, only those probe pairs with distance less than a threshold θ can be possibly aligned. Proposition 2 states that the number of such pairs is upper bounded by $\frac{\theta m}{\delta}$, where δ is the minimal distance between the adjacent probe locations in V. Thus, the dynamic programming only needs to explore a linear number of pairs defined as a function of θ and δ. The complete proof of the two propositions are given in Supplementary Section 2.

3 EXPERIMENTS

We evaluated both the probe alignment kernel and the MPA algorithm with experiments on three bladder cancer datasets as well as simulations. Support vector machines (SVMs) were used as the classifier in all the classification tasks. Both linear and RBF kernels were tested for the interpolation method. In all experiments, we tested SVM parameter $C=\{10^{-4},10^{-3},\ldots,10^{4}\}$ and $RBF\_\sigma=\{10^{-3},10^{-2},\ldots,10^{3}\}$.

3.1 Simulations

In the simulations, we first generated 10 000 possible locations for probe sampling with a Markov model. We assume that locations are from two types of chromosome regions: gene-rich and non-coding regions. The probe density in gene-rich regions is higher than that in non-coding regions. Specifically, the distance between adjacent locations is 1 in gene-rich regions and 10 in non-coding regions. The Markov model takes two states {'gene-rich region', 'non-coding region'} with a transition probability 0.9 for staying at a state and 0.1 for jumping between the states. Thus, continuous gene-rich or non-coding regions with variable lengths will be generated by the Markov model. We randomly generated 50 samples of probe series in the case group and another 50 in the control group by random sampling from the 10 000 locations in each experiment. We simulated the DNA amplification and deletion events in the case samples for two test scenarios. In the first scenario, we randomly selected 20 regions (10 amplifications and 10 deletions), each of which consists of 10 consecutive locations out of the 10000 locations as discriminant CNVs on the chromosome. In the second scenario, we randomly selected 20 regions with variable numbers of locations between 1 and 20. This strategy generates CNV regions with significantly different lengths to mimic short chunks as well as large chunks of DNA amplification/deletions. The feature value of a probe that is not in a CNV region is generated from a normal distribution $N(0,0.5^2)$. The rule implies that in normal individuals, a DNA amplification or deletion is a relatively rare event. A probe in the CNV regions in a case sample will take a value from a normal distribution $N(1/-1,0.5^2)$ to indicate a measured CNV value at the locations in the amplification/deletion regions.

We compared the standard interpolation with the alignment kernel using the cost function defined in Equation (1) and a small gap $g=0.001$. In the standard interpolation, a linear interpolation maps the missing positions in each series. The feature value at an interpolated position is assigned the distance-weighted average of the two nearest positions. After interpolation, each sample will have a feature vector of the same dimension, and standard classifiers such as SVM can be used for classification. We randomly generated datasets and tested the classification accuracy with a 5-fold cross-validation. In each fold, another 4-fold cross-validation on the training set is used to tune parameters for each algorithm. The average accuracy of each algorithm is reported in Table 1. The alignment kernel clearly outperformed the interpolation method. When probes are sparse, the improvement is up to 12%. When there are more probes available, the improvement is smaller. Especially, when 1000 probes are sampled, the improvement is only 3%. In the experiments with variable number of probes per sample or with variable number of locations per amplification/deletion region, the alignment kernel also significantly outperformed the interpolation. The experiments proved that the alignment kernel could handle sparse probe series with varying numbers of probes or varying lengths of amplifications/deletions well for classification.

To compare the multiple alignment algorithm with the interpolation method, we also tested alignment of multiple probe series for identifying the true common CNV regions. The multiple alignment algorithm used the scoring function defined by Equation (2) with $\tau=\frac{\sigma}{10}$. For better visualization, we only report the results on small datasets without distinguishing gene-rich regions and non-coding regions although similar results were observed on larger datasets. We generated five samples with 100 locations, within which there are two amplification regions and two deletion regions. In the example in Figure 2A, each region contains five consecutive locations. In the example in Figure 2B, each region contains 1–10 consecutive locations. We randomly sampled 10 features for both cases. From the two examples, the result of multiple alignment is clearly more accurate in detecting the exact locations of amplifications and deletions. Since the interpolation approach propagates the information from a probe to its interpolated neighbors, this propagation-based assumption often results in blurry boundaries and fails to distinguish the close-by CNV events, as visualized in the examples in Figure 2 and Supplementary Figure 6. The results suggest that the multiple alignment algorithm is more robust to tolerate probe sparsity and noise in the data.

3.2 Bladder cancer datasets

We collected three different arrayCGH datasets generated for studying bladder cancer. The first dataset (D_1) introduced by Blaveri *et al.* (2005) was generated with a HumanArray 2.0 array consisting

Table 1. Classification accuracy on artificial datasets. SVMs were used as the classifier with the kernels

Method	Uniform probe density					Varying probe density		
	$n=100$	$n=500$	$n=1000$	$n=U(1-1000)$	$n=N(500,100^2)$	$n=100$	$n=500$	$n=1000$
Linear kernel	0.502	0.759	0.922	0.701	0.764	0.536	0.814	0.963
RBF kernel	0.519	0.719	0.863	0.659	0.716	0.535	0.753	0.905
Align kernel	**0.597**	**0.875**	**0.950**	**0.811**	**0.857**	**0.601**	0.85	0.968

The accuracies are averages of 10 random experiments for each case. n is the number of probes in each series. The probes are a random subset of the 10 000 locations generated for each sample in three ways: (i) constant numbers (from 100 to 1000) of features were extracted out of the 10 000 features for a sample such that each sample has the same number of features but at different locations; (ii) different numbers of features were extracted for each sample from a discrete uniform distribution on [1, 1000]; and (iii) different numbers of features for each sample were generated from a normal distribution $N(500, 100^2)$ (rounded to positive integers). The scores that are statistically significantly larger than the other compared average scores are bold.

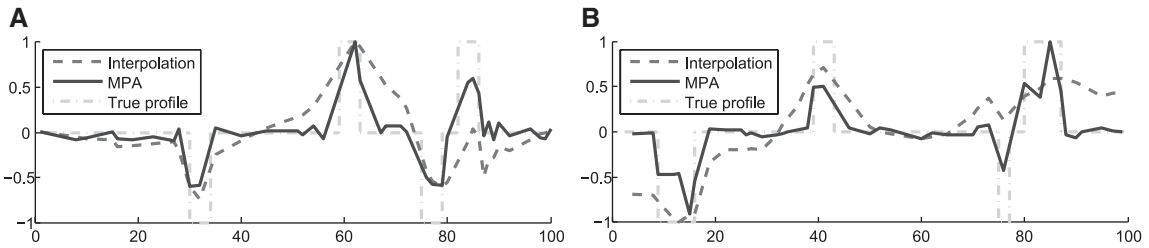

Fig. 2. Comparison between MPA and interpolation in simulation. CNV profiles generated by multiple alignment algorithm (MPA) and interpolations are compared with the four true common CNV regions plotted in the step function. All the values in the plots were rescaled to the range $[-1, 1]$ for better visualization. (**A**) Example of uniform probe density. (**B**) Example of varying probe density.

of 2464 probes at 1.5 Mb resolution. The second dataset (D_2) introduced by Stransky *et al.* (2006) was also generated with a HumanArray 2.0 array but consisting of 2385 probes at 1.3 MB resolution. The third dataset D_3 introduced by Heidenblad *et al.* (2008) was generated with a high-resolution tiling BAC Re-Array set 1.0 containing 36 288 BAC probes. The datasets were post-processed by the authors and the clones from sexual chromosomes were not included in the study because they are not comparable between male and female samples. After the pruning, dataset D_1 contains 98 samples and 2142 probes, dataset D_2 contains 57 samples and 2308 probes, and dataset D_3 contains 38 samples and 24 384 probes. We tested classification of patient samples by tumor grades in the three datasets. The two classes are distinguished by lower tumor grades or higher tumor grades. Specifically, we considered 'Low' versus 'High' in D_1, and '≤G2' versus '>G2' in D_2 and D_3 (Rapaport *et al.*, 2008; Tian *et al.*, 2009).

3.2.1 Parameter selection The σ parameter is the normalization term of the probe distances in Equations (1) and (2), and the gap penalty g defines the scoring scale of a gap in the alignment. We set σ to be the average distance between all pairs of probes on a chromosome, based on the statistics on each chromosome in Supplementary Table 2. Note that different σs were used for probes on different chromosomes. In the three datasets, the σs for the 22 chromosomes are all in the order of 10^4–10^5. Given the estimated σ from the data, an additional cross-validation on the training set is performed to choose the gap penalty g in $\{10^{-5}, \ldots, 10^{-1}\}$. In multiple alignment, σ can be chosen in the same way, but the g parameter is actually decided by the minimal allowed matching distance τ [Equation (2)]. Since g will only take values in the range $[e^{-1}, 1]$ if $\tau < \sigma$, empirically any $\tau < \sigma$ can give reasonable multiple alignment on both artificial and real datasets. From the definition $g = e^{-\frac{\tau}{\sigma}}$, we can expect the results of multiple alignment will not be sensitive to τ as long as $\frac{\sigma}{10} < \tau < \sigma$. Thus, in the multiple alignments, we set $\tau = \frac{\sigma}{10}$.

3.2.2 Classification of patient samples To evaluate the classification performance of the alignment kernel, we performed a special *cross-dataset validation*. We first selected a target dataset to run 5-fold cross-validation. In the experiment on each fold, we also used the samples in the other two datasets as additional training samples. We performed the cross-dataset validation for D_1, D_2 and D_3, each with 50 times of random 5-fold cross-validation. One additional baseline is the raw kernel function defined by Liu *et al.* (2008), which defines pairwise relations between CNV states as a new variation of linear kernel. Note that the experiments of all the compared methods were on the same 5 folds of each dataset. In the experiment on each fold, another 4-fold cross-validation on the training set was performed to tune parameters for each algorithm, specifically, the SVM C parameter and the gap parameter g for probe alignment, the SVM C parameter for linear kernel and raw kernel, and the SVM C parameter and the RBF_σ for RBF kernel.

Two categories of comparisons were performed. First, we tested SVMs used with linear kernel, RBF kernel and raw kernel on the three datasets independently but with the alignment kernel using additional training data from the other two datasets. These experiments were purposed to show that integrating samples from other datasets in the training set can generate more accurate classifiers. The results are reported in Table 2. Clearly, using the additional data, the alignment kernel significantly improved the

classification accuracy on the three datasets. Next, the alignment kernel was compared with the other kernels on the interpolated data with the interpolated features on multiple datasets in all possible combinations. The average accuracy of all the methods are listed in Table 3. The alignment kernel outperformed or tied with the best of linear kernel, RBF kernel and raw kernel on the interpolated data in almost all the cases except it is the second best when tested with D_3 as target and D_1 for additional training.

3.2.3 Detecting common CNV regions To evaluate how well the three datasets can be integrated to detect the common CNV regions in bladder cancer tumors, we applied the multiple alignment algorithm to align the samples with higher tumor grades in the three datasets, and compared the alignment result with the interpolation result. Since all the samples in the same dataset have identical probe locations, we first calculated the average log-ratio intensities of the three datasets to get three consensus CNV probe series. The multiple alignment method was then used to produce an *MPA profile* of the three consensus probe series. The probe intensities in the profile are rescaled to be in the range [−1, 1]. We then report the probe positions with an absolute intensity larger than 0.4 and identified 49 regions with an average length of 20 000 kb. We also used linear interpolation to get an average of all samples and also rescaled the value to be in range [−1, 1]. To make a fair comparison with alignment method, we reported 708 center regions in the interpolation profile, each with absolute intensity values larger than 0.8. The positions that are centered at the 708 regions with length 10 000 kb were taken as the common CNV regions. The results

Table 2. Improvement of bladder cancer classification from data integration with probe alignment kernel

Method	D_1	D_2	D_3
Linear kernel	0.808	0.612	0.835
RBF kernel	0.824	0.603	0.806
Raw kernel	0.812	0.642	0.846
Alignment kernel	**0.870**	**0.797**	**0.883**

The accuracies are averages of 50 random 5-fold cross-validations.

Table 3. Comparison of the probe alignment kernel and interpolation in cross-dataset validation

Additional training	Target dataset	Interpolation			Alignment kernel
		Linear	RBF	Raw	
D_2	D_1	0.805	0.817	0.816	**0.829**
D_3	D_1	0.812	0.833	0.828	**0.868**
D_1	D_2	0.635	0.707	0.659	**0.791**
D_3	D_2	0.666	0.755	0.674	0.766
D_1	D_3	0.819	0.785	**0.893**	0.847
D_2	D_3	0.839	0.812	0.852	0.848
D_2, D_3	D_1	0.813	0.842	0.797	**0.870**
D_1, D_3	D_2	0.668	0.785	0.664	**0.797**
D_1, D_2	D_3	0.812	0.834	0.881	0.883

The accuracies are averages of 50 random 5-fold cross-validations. The second column shows the target dataset for 5-fold cross-validation. The first column shows the datasets used as additional training samples.

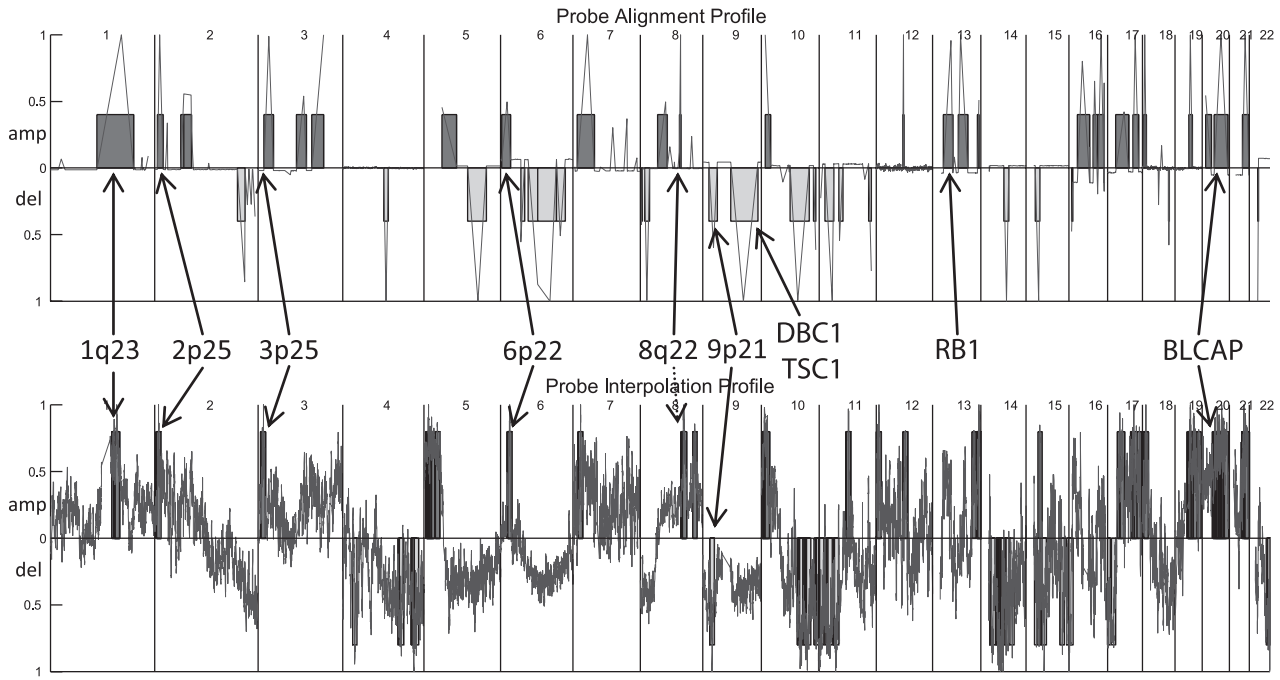

Fig. 3. Comparison of MPA profile and interpolation profile. The MPA profile is above the interpolation profile. The common DNA amplification and deletion regions are plotted with blocks above or under the *x*-axis, respectively. The probes in the profiles are ordered by their locations on chromosomes (from chromosome 1 to chromosome 22) and the corresponding intensities are plotted by the curves. The vertical lines represent the chromosome separations. The locations of the known cancer-related CNVs and the four bladder cancer genes (DBC1, TSC1, RB1 and BLCAP) are marked with arrows.

of MPA and interpolation are compared in Figure 3. Many of the common CNV regions identified by the two methods are similar, but the alignment profile is smoother. The interpolation method reported many more short fragments, many of which are false positives. One remarkable finding is that the MPA captured four known bladder cancer genes in the common CNV regions, while the interpolation only identified one.

The MPA detected four bladder cancer genes, DBC1, TSC1, RB1 and BLCAP, and several common amplification/deletion regions associated with bladder cancer. DBC1 (deleted in bladder cancer 1) is a gene that shows loss of heterozygosity in some bladder tumors. TSC1's somatic mutations are associated with the tumor cases with one copy of the TSC1 gene missing due to a deletion in chromosome 9. The two genes were identified at the end of the second deletion regions in the alignment profile on chromosome 9. RB1 with mutations, believed to contribute to the development of bladder cancer, is identified at the first amplification region on chromosome 13. BLCAP (bladder cancer-associated protein), believed to be involved in the cancercinogenesis, is identified at the second amplification region at chromosome 20. Besides these bladder cancer genes, the MPA also identified some common amplification and deletion regions associated with bladder cancer. For example, we identified an amplification event at location 6p22 with a main target gene E2F3 (Oeggerli *et al.*, 2006). Deletions of part of or all chromosome 9 are common events in bladder tumors and we identified these events including a known location 9p21. The results are compared with previous studies on the datasets (Blaveri *et al.*, 2005; Heidenblad *et al.*, 2008) in Supplementary Table 3. The MPA can retrieve most findings from these independent studies, which implies that probe alignment successfully integrated information from the three datasets. The multiple alignment also missed some amplification/deletion regions, many of which might be false positives that were manually identified in a very small number of samples.

4 DISCUSSIONS

Integration of diverse genomic datasets has become one of the central problems in cancer genomics. Most of the previous work on arrayCGH data analysis focused only on segmentation of the probe series for deriving real CNV events. In this article, we introduce a general probe alignment framework to integrate arrayCGH datasets generated on different platforms. We demonstrated with experiments that the probe alignment-based approaches have a good potential to generate significantly improved classification performance and detect more accurate common CNVs. The results suggest that these approaches are powerful tools for integrative studies of multiple arrayCGH datasets. There are three technical issues in applying probe alignment: fast implementation, parameter tuning and kernel selection.

- We tested the fast probe alignment described in Section 2.4 to show that only a small fraction of the entries in the dynamic programming matrix needs to be computed. Although the probe locations and log-ratios are highly variable in different platforms, we empirically observed that the time complexity is indeed close to linear in our experiments (Supplementary Fig. 1). However, it is also possible that a larger gap penalty is needed to reduce the time complexity significantly, which might lead to worse classification results. In this case, approximation is necessary to restrict the alignment to a smaller number of probe pairs to produce sub-optimal alignment scores for classification.
- Intuitively, a good σ should be in the same scale of the average probe distances as estimated in Supplementary Table 2 to distinguish the difference in the distances between matched probes. Small positive gs will result in identical alignment and similar good classification results with less impact from the gaps. We verified the intuitions by additional classification experiments in Supplementary Fig. 2. For a more rigorous treatment of the problem, based on a comprehensive analysis of the datasets (Supplementary Figs 3–5 and Table 2), we suggest a strategy that estimates the σ as the average probe distance and selects a gap penalty by a cross-validation in the training set, as a robust strategy to choose good parameters for SVM classification. For the multiple alignment case, we suggest the same strategy for choosing σ and taking a gap penalty g that is decided by the allowable distance between the matched probes for generating reasonable multiple alignment. In practice, both strategies worked well.
- Finally, the probe alignment kernel is not guaranteed PSD. Besides adding a positive constant on the diagonal, another alternative that we tested is to use the nearest PSD matrix of the alignment-score matrix as the kernel matrix (Higham, 1988). However, we observed that the nearest PSD matrix is not a good kernel in the experiments (Supplementary Table 1).

A promising future direction is to combine segmentation approaches with alignment. The segmentation approaches detect the actual CNV intervals in each probe series and the alignment approach can be modified to align the intervals. We performed an initial segmentation analysis on the three bladder cancer datasets with the segmentation method proposed in Olshen *et al.* (2004). At the suggested significance level, the segmentation detected very few intervals of amplifications and deletions that can be used for alignment. Thus, one difficulty is how to choose the parameters such as window length or significance levels for getting sensible segmentation results for alignment. In addition, aligning the detected CNV intervals in segmentation requires development of a new alignment approach to compare CNV intervals. Thus, the alignment approach in this article is not directly applicable. Another possible extension is to distinguish regions with different probe density (gene-rich versus non-coding) with different normalization for further improvement of classification or common CNV detection. Certainly, this extension requires more sophisticated treatment of the alignment functions, which will introduce additional complexity to the problem.

ACKNOWLEDGMENTS

This work was carried out in part using computing resources at the University of Minnesota Supercomputing Institute.

Funding: Biomedical Informatics and Computational Biology Seed Grant from University of Minnesota Rochester.

Conflict of Interest: none declared.

REFERENCES

Aach,J. and Church,G. (2001) Aligning gene expression time series with time warping algorithms. *Bioinformatics*, **17**, 495–508.

Altschul,S. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Blaveri,E. *et al.* (2005) Bladder cancer stage and outcome by array-based comparative genomic hybridization. *Clin. Cancer Res.*, **11** (Pt 1), 7012–7022.

Carter,N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.*, **39** (Suppl. 7), S16–S21.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge.

Feuk,L. *et al.* (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.

Guha,S. *et al.* (2008) Bayesian hidden Markov modeling of array CGH data. *J. Am. Stat. Assoc.*, **103**, 485–497.

Heidenblad,M. *et al.* (2008) Tiling resolution array CGH and high density expression profiling of urothelial carcinomas delineate genomic amplicons and candidate target genes specific for advanced tumors. *BMC Med. Genomics*, **1**, Article 3.

Higham,N. (1988) Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl.*, **103**, 103–118.

Liao,L. and Noble,W. (2003) Combining pairwise-sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Biol.*, **10**, 857–868.

Liu,J. *et al.* (2008) Classification and feature selection algorithms for multi-class CGH data. *Bioinformatics*, **24**, I86–I95.

Oeggerli,M. *et al.* (2006) E2F3 is the main target gene of the 6p22 amplicon with high specificity for human bladder cancer. *Oncogene*, **25**, 6538–6543.

Olshen,A. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Rapaport,F. *et al.* (2008) Classification of arrayCGH data using fused SVM. *Bioinformatics*, **24**, I375–I382.

Redon,R. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Shlien,A. and Malkin,D. (2009) Copy number variations and cancer. *Genome Med.*, **1**, 62.

Stransky,N. *et al.* (2006) Regional copy number-independent deregulation of transcription in cancer. *Nat. Genet.*, **38**, 1386–1396.

Sykes,N.H. *et al.* (2009) Copy number variation and association analysis of SHANK3 as a candidate gene for autism in the IMGSAC collection. *Eur. J. Hum. Genet.*, **17**, 1347–1353.

Thompson,J. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Tian,Z. *et al.* (2009) A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge. *Bioinformatics*, **25**, 2831–2838.