BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 19 2005, pages 3763–3770
doi:10.1093/bioinformatics/bti611

Genetics and population analysis

Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data

Weil R. Lai[1], Mark D. Johnson[2], Raju Kucherlapati[1] and Peter J. Park[1,3,*]

[1]Harvard-Partners Center for Genetics and Genomics, 77 Avenue Louis Pasteur, Boston, MA 02115, USA, [2]Department of Neurological Surgery, Brigham and Women's Hospital and Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA and [3]Children's Hospital Informatics Program, 300 Longwood Ave, Boston, MA 02115, USA

Received on June 17, 2005; revised on July 12, 2005; accepted on August 2, 2005
Advance Access publication August 4, 2005

ABSTRACT

Motivation: Array Comparative Genomic Hybridization (CGH) can reveal chromosomal aberrations in the genomic DNA. These amplifications and deletions at the DNA level are important in the pathogenesis of cancer and other diseases. While a large number of approaches have been proposed for analyzing the large array CGH datasets, the relative merits of these methods in practice are not clear.

Results: We compare 11 different algorithms for analyzing array CGH data. These include both segment detection methods and smoothing methods, based on diverse techniques such as mixture models, Hidden Markov Models, maximum likelihood, regression, wavelets and genetic algorithms. We compute the Receiver Operating Characteristic (ROC) curves using simulated data to quantify sensitivity and specificity for various levels of signal-to-noise ratio and different sizes of abnormalities. We also characterize their performance on chromosomal regions of interest in a real dataset obtained from patients with Glioblastoma Multiforme. While comparisons of this type are difficult due to possibly sub-optimal choice of parameters in the methods, they nevertheless reveal general characteristics that are helpful to the biological investigator.

Contact: peter_park@harvard.edu

INTRODUCTION

Locating chromosomal aberrations in genomic DNA samples is an important step in understanding the pathogenesis of many diseases. This is especially true in cancer, and an enormous amount of effort and resources has been dedicated to the detailed characterization of the chromosomal abnormalities in the development and progression of various cancers. Amplification or deletion of chromosomal segments can lead to abnormal mRNA transcript levels and results in the malfunctioning of cellular processes.

Array comparative genomic hybridization (CGH) is a technique for measuring such changes (Solinas-Toldo *et al*., 1997; Pinkel *et al*., 1998). See Pinkel and Albertson (2005) for a review. The main difference between array CGH and mRNA expression profiling is that genomic DNA rather than mRNA transcripts are hybridized in array CGH. As the resolution of the arrays has improved over the years, array CGH has become a powerful tool. As a high-throughput technique, it offers many advantages over other cytogenetic techniques such as fluorescence *in situ* hybridization (FISH). While early experimental techniques were only able to detect chromosomal changes at the whole chromosomal or whole arm level, the CGH arrays using BAC (Bacterial Artificial Chromosome) clones have been widely used subsequently, with the resolution on the order of 1 Mb (Pinkel *et al*., 1998). These arrays generally contain many regions with known oncogenes and tumor suppressor genes, and can be iteratively designed in a locus-specific manner to identify candidate genes in a small region. More recently, cDNA and oligonucleotide arrays have become popular for CGH (Pollack *et al*., 1999; Brennan *et al*., 2004). The shorter probes on these new arrays may not be as robust as BACs for large segments, but they offer much higher resolution (in the order of 50–100 kb). In particular, oligonucleotide arrays allow design flexibility and greater coverage, and they appear to provide sufficient sensitivity (Brennan *et al*., 2004). Tiling or custom arrays are also available now for even finer resolution of specific regions and allow the detection of micro-amplifications and deletions (Lucito *et al*., 2003; Ishkanian *et al*., 2004).

The resultant high-throughput array CGH data have prompted the development of various algorithms for data analysis, as briefly reviewed in the next section. However, while there have been numerous publications introducing new methods, the relative strengths and weaknesses of these methods are difficult to discern, due to the complexity of the algorithms and the lack of software with visualization tools. This problem is exacerbated by nondescript titles and abstracts of the articles and their lack of extensive performance comparisons to existing methods. This is especially true from the perspective of the biologist who must choose an algorithm for the dataset of interest. The purpose of this paper is to compare the algorithms that have been published so that the user can quickly gain an overview of the array CGH algorithms and their performance. Both simulated data and real data obtained from glioblastoma samples are used for evaluating the algorithms. The methods evaluated in this paper are listed in Table 1.

METHODS

Basic issues and algorithms

Array CGH data consist of the log-ratios of normalized intensities from disease vs control samples, indexed by the physical location of the probes on the genome. The goal is to identify regions of concentrated high or low log-ratios.

*To whom correspondence should be addressed.

Table 1. List of algorithms tested in this paper.

Name	Reference	Method	Software	Type
CGHseg	Picard *et al*. (2005)	CGH Segmentation	CGHseg, Nov, 2004 (MATLAB)	E
Quantreg	Eilers and de Menezes (2005)	Quantile Smoothing	quantreg, v3.76 (R)*	S
CLAC	Wang *et al*. (2005)	Clustering Along Chromosomes	CLAC, v0.1-1 (R)	S, E
GLAD	Hupe *et al*. (2004)	Adaptive Weights Smoothing	GLAD, v1.0.2 (R)	S, E
CBS	Olshen *et al*. (2004)	Circular Binary Segmentation	DNAcopy, v1.1.1 (R)	E
HMM	Fridlyand *et al*. (2004)	Hidden Markov Model	aCGH, v1.1.4 (R)	E
Wavelet	Hsu *et al*. (2005)	Maximal Overlap Discrete Wavelet Transform	waveslim, v1.4 (R)*	S
Lowess		Locally Weighted Regression	stats, v2.0.1 (R)*	S
ChARM	Myers *et al*. (2004)	Chromosomal Aberration Region Miner	ChARM, v1.6 (JAVA)	S, E
GA	Jong *et al*. (2003)	Genetic Local Search	aCGHSmooth, Nov, 2004 (exec)	E
ACE	Lingjaerde *et al*. (2005)	Analysis of Copy Errors	CGH-Explorer, v2.3 (JAVA)	S, E

For the last column, 'S' and 'E' indicate that the algorithm has a step for smoothing and estimation, respectively. Three methods (Quantreg, Wavelet and Lowess) are for smoothing only. Some methods or packages did not have specific names; others had names that are too generic. We have created short abbreviations in such cases [e.g. we have called the method in Picard *et al*. (2005) based on the name of their downloadable file]. These names are used in the subsequent figures. *indicates those using existing R packages: Quantreg and Wavelet methods were implemented by us based on the descriptions given in the papers; Lowess is our implementation using the existing R function. CGHseg was ported to R from MATLAB by us. A list of websites for these packages can be found in the Supplementary Material available at http://www.chip.org/∼ppark/arrayCGH_comparison.

In general, these regions of interest can be very small; some microdeletions may only contain a single probe. Because attempting to identify such small regions can result in too many false positives, information from consecutive probes are used to identify larger regions with more confidence.

The first analytical methods were simple yet intuitive and often effective, involving smoothing of the ratio profiles and applying a reasonable threshold to determine if the average ratio over a potential region signified an amplification or a deletion. For instance, a moving average was used to process the ratios, and a 'normal versus normal' hybridization was used to compute a threshold level (Pollack *et al*., 2002). In another study, a simple maximum likelihood method was used to fit a mixture of three Gaussian distributions corresponding to gain, loss and normal regions (Hodgson *et al*., 2001).

Broadly, there are two estimation problems. One is to infer the number and statistical significance of the alterations; the other is to locate their boundaries accurately. The many available methods differ in the ways in which each part is modeled and the two are combined. In general, the formulation of a model-based method presumes a sequence of piecewise constant segments as a function of various parameters such as the number of breakpoints, their locations and the mean/variance of the distributions for each segment. Then the maximization of a function, typically a log-likelihood, is used to estimate the model parameters from the data. In the likelihood, a penalty term for the number of segments is often included to avoid too fine a partition, which tends to increase the likelihood. Models differ in their distributional assumptions and the incorporation of penalty terms.

Subsequently, more complicated methods for denoising and estimating the spatial dependence were derived. Genomic amplifications and deletions are assumed to cover multiple probes in general, and an effective incorporation of this spatial structure is a key component in any algorithm. For instance, a quantile smoothing method based on the minimization of errors in L_1 norm (sum of absolute errors) rather than L_2 norm (sum of squared errors) is shown to give sharper boundaries between segments (Eilers and de Menezes, 2005). Another promising smoothing algorithm is a denoising by wavelets (Hsu *et al*., 2005), a nonparametric technique that appears to handle abrupt changes in the profiles well. A simple and more common approach is based on robust locally weighted regression and smoothing scatterplots (lowess), introduced in Cleveland (1979). This has been used previously in other works such as in Beheshti *et al*. (2003).

In Olshen and Venkatraman (2002) and Olshen *et al*. (2004) the binary segmentation method (Sen and Srivastava, 1975) is modified to allow splits into either two or three segments. In this algorithm, termed Circular Binary Segmentation (CBS), the maximum of a likelihood ratio statistic is used recursively to detect narrower segments of aberration. In Jong *et al*. (2003, 2004), a genetic search algorithm is used to maximize a likelihood with a penalty term containing the number of breakpoints. In Hupe *et al*. (2004), a more complex likelihood function with weights determined adaptively is used to solve the estimation problem locally based on data smoothed by the Adaptive Weights Smoothing procedure (Polzehl and Spokoiny, 2000). A likelihood method with a different penalty function is used in Picard *et al*. (2005) for the number of segments to avoid underestimation on them. It is pointed out that a distribution assumption can have an important consequence in a model: a homogeneous variance assumption among different regions, for example, tends to lead to a more segmented profile in order to satisfy the variance assumption. In Daruwala *et al*. (2004), a Poisson distribution is used to model the number of segments and this is incorporated as an additional component in the likelihood. (This last method was not available publicly.)

In a more local approach (Myers *et al*., 2004) an edge filter is used to detect the approximate location of edges, and an EM algorithm is used to place them more precisely. In Lingjaerde *et al*. (2005), a simple smoothing is done using signs of neighbors, and significance is determined by comparing both the width and height of the observed segments with their joint null distribution. A dynamic programming approach is used in Autio *et al*. (2003), but this was not part of our study because the associated MATLAB package was difficult to port to our platform.

A different kind of modeling approach involves the Hidden Markov models (HMMs), in which the underlying copy numbers are the hidden states with certain transition probabilities (Snijders *et al*., 2003; Sebat *et al*., 2004; Fridlyand *et al*., 2004). In Wang *et al*. (2005), a simple but effective method based on hierarchical clustering along the chromosomes is used to identify regions of interest and the False Discovery Rate (FDR) is used as a selection criterion.

Evaluation method

Evaluation of the relative performance of these methods is complicated by several problems. One difficulty is that the goals of different algorithms are not the same. For example, those with an emphasis on the smoothing part may simply return the log-ratios without determining which ones are significant. More comprehensive methods may return the coordinates of only the statistically significant segments with or without the estimated average log-ratio per segment. One may require a 'normal versus normal' sample as a control

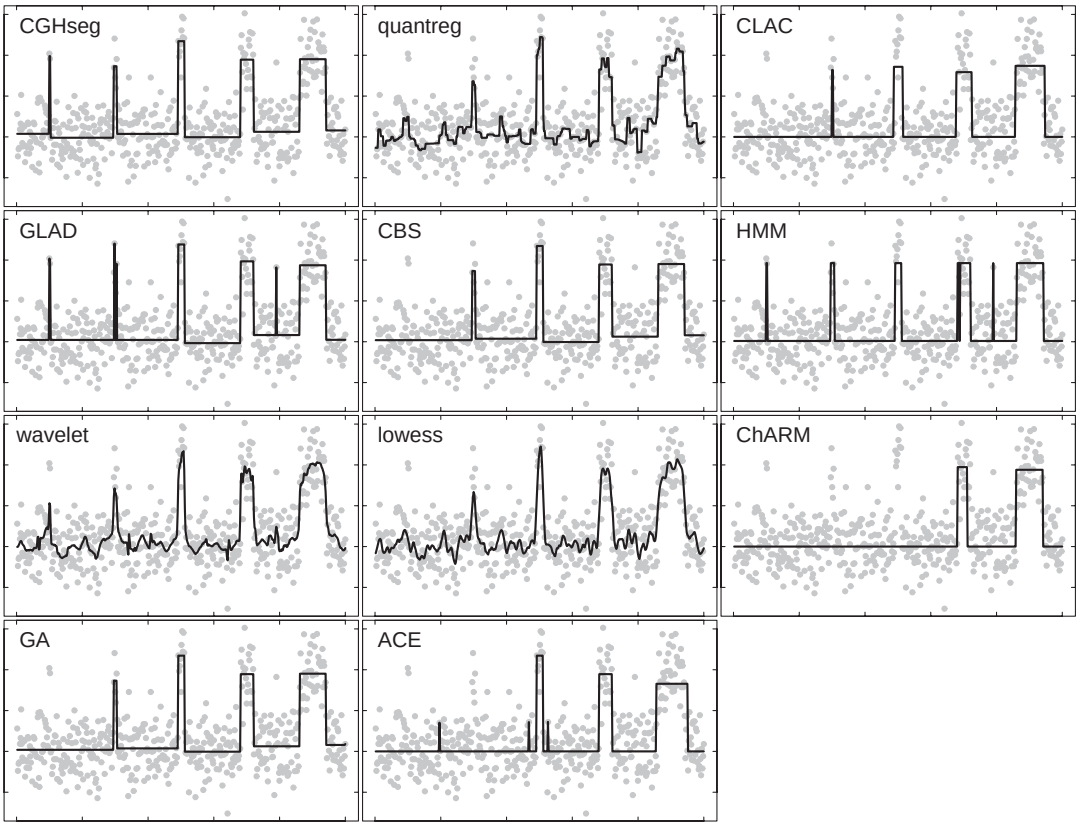

Fig. 1. Array-CGH algorithms on simulated aberrations of increasing width. Illustrated here as an example are the signal profiles consisting of five aberrations of 2, 5, 10, 20 and 40 probes long with an amplitude of 1. Gaussian noise $N(0, .25^2)$ was added onto the signal profile to generate the simulated data. Default settings for the algorithms were used when available; otherwise, appropriate parameters were selected or computed based on the program documentation and related papers.

while another may not. When necessary, we simulated the control samples by Gaussian noise with zero mean and a variance estimated from the tested data using the median absolute deviation.

In terms of implementation, the primary issue was that not every algorithm was implemented in a publicly available software. All our calculations were carried out in the statistical language R (R Development Core Team, 2004) http://www.R-project.org, since this was the dominant platform for the software packages. But other algorithms were implemented in MATLAB, a JAVA application, or an executable file. When the code could be ported easily from MATLAB to R, such as the algorithm in Picard *et al*. (2005), this was carried out. When a program allowed a relatively simple interface, as was the case with the algorithms in Jong *et al*. (2004), Myers *et al*. (2004) and Lingjaerde *et al*. (2005), it was used for computations.

The most appropriate way to compare these algorithms on simulated data was to calculate the Receiver Operating Characteristic (ROC) curves. Because different algorithms were tuned at different sensitivity levels, it was important to examine the trade-off between sensitivity and specificity in each case. This approach adjusts for the differences arising from identifying different numbers of segments in each algorithm. We have used the default parameters for each algorithm, as most users will be doing. If no default parameters were available, we used the steps suggested in the program documentation or the papers describing the method to select the parameter values. We have not attempted to adjust the parameters to improve the performance, due to the large number of algorithms and the large number of scenarios under which they were tested. This issue is further discussed in the section Discussion. The general properties are nonetheless apparent, even if the parameters were suboptimal.

RESULTS

First, we tested the algorithms on simulated data of various abnormality widths and noise levels. To generate ROC curves corresponding to a particular aberration width and noise level, we calculated the true positive rates (TPR) and the false positive rates (FPR) as we varied the threshold for determining an aberration. We also tested the algorithms on cDNA microarray data containing measurements from 26 different primary Glioblastoma Multiforme (GBM) tumors (Bredel *et al*., 2005).

Simulated data

We calculated the ROC profiles of each algorithm for aberration widths of 5, 10, 20 and 40 probes, and signal-to-noise ratios (SNR) of 1, 2, 3 and 4. SNR was defined as the mean magnitude of the aberration (i.e. signal) divided by the standard deviation of the superimposed Gaussian noise. Figure 1 illustrates the kind of profiles examined in this simulation. For each aberration width and SNR, we generated 100 artificial chromosomes, each consisting of 100 probes and with the square-wave signal profile added to the center of the chromosome. The performance of the algorithms for the aberrations at the boundaries was not examined here.

TPR was defined as the number of probes inside the aberration whose fitted values are above the threshold level divided by the number of probes in the aberration. FPR was defined as the number

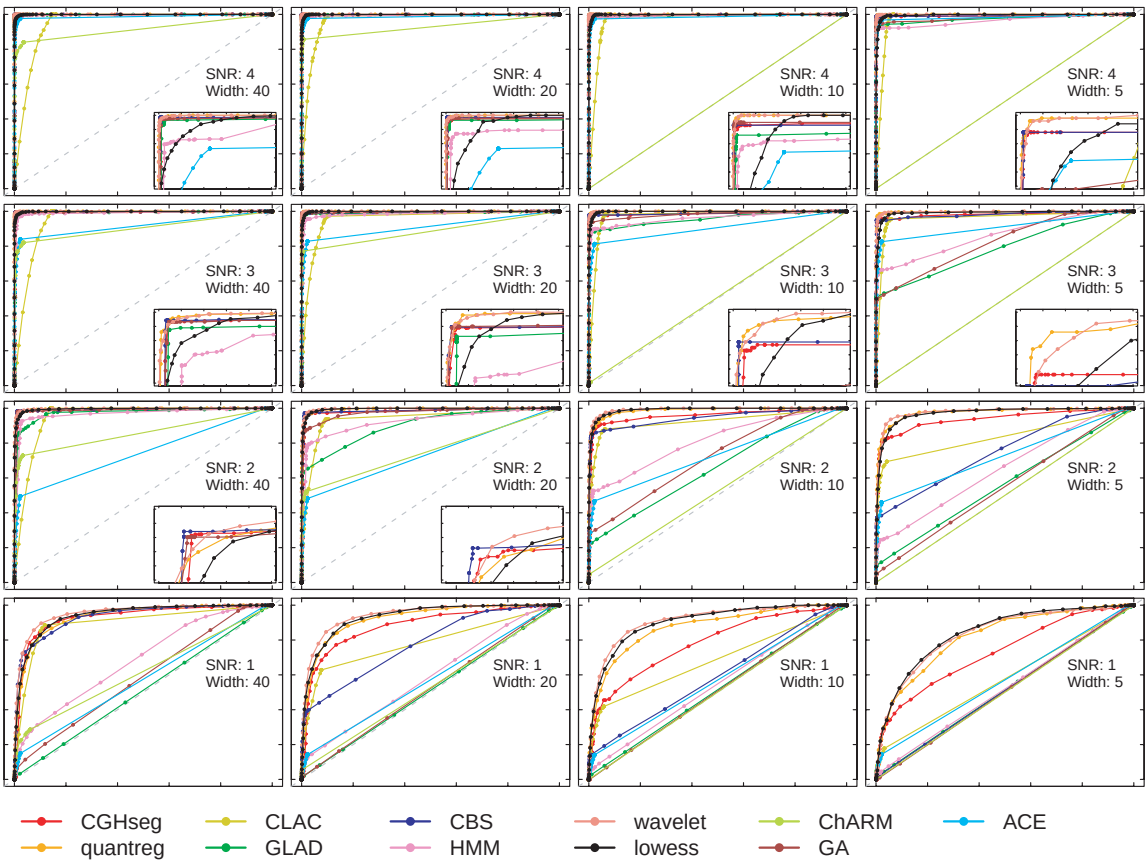

Fig. 2. Receiver operating characteristic (ROC) curves for array CGH algorithms measured at different aberration widths and signal-to-noise ratios (SNR). The x-axis is the false positive rate and the y-axis is the true positive rate. Red is CGHseg (Picard *et al*., 2005), orange is quantreg (Eilers and de Menezes, 2005), dark yellow is CLAC (Wang *et al*., 2005), green is GLAD (Hupe *et al*., 2004), blue is CBS (Olshen *et al*., 2004), violet is HMM (Fridlyand *et al*., 2004), salmon is wavelet (Hsu *et al*., 2005), black is lowess, light green is ChARM (Myers *et al*., 2004), brown is GA (Jong *et al*., 2003) and cyan is ACE (Lingjaerde *et al*., 2005). The curves were generated by measuring the true and false positive rates on simulated data at different threshold levels.

of probes outside the aberration whose fitted values are above the threshold level divided by the total number of probes outside the aberration. In order to compute the ROC curve, we varied the threshold value for aberration from the minimum log-ratio value to the maximum. (This is equivalent to moving the x-axis cutoff value in a mixture distribution.) Each threshold value results in a TPR and a FPR, represented by a point on the ROC curve. A set of TPRs and FPRs were then plotted to reveal the algorithm's ROC profile for the particular aberration width and SNR (see Fig. 2). We also computed confidence intervals around each ROC curve (data not shown) but significant differences among the methods did not appear to exist.

We note that TPR and FPR are informative in understanding how an algorithm performs in estimating the boundary of the altered region. When the algorithm over-estimates the boundary, FPR increases while TPR remains fixed; when it under-estimates the boundary, TPR decreases while FPR remains fixed. We also note that FPR estimates depend on the size of the aberration relative to that of the chromosome. Therefore, FPR should be used only for measuring relative performance among different methods given a fixed aberration size.

The default parameters in the software were used except in the following cases where the parameters had to be chosen by the user. For quantile smoothing, we followed the suggestion of Eilers and de Menezes (2005) to use 2-fold cross-validation to estimate the value of λ that minimizes the overall penalty term; this gave us a λ of 1.5. For the wavelet denoising algorithm, we chose soft Stein's Unbiased Risk Estimate thresholding with a maximum wavelet coefficient level of 3 based on results given in their paper. For lowess, we used a smoothing window of 10 probes and defined the smoothing span as the size of the smoothing window divided by the number of non-missing log-ratios in the chromosome. For the genetic algorithm (Jong *et al*., 2003), we deselected the option to filter out log-ratios with a threshold value of 0.75 so that their program will consider all data points in the analysis. For Analysis of Copy Errors (Lingjaerde *et al*., 2005) we chose results for the estimated false discovery rate closest to 0.001.

As shown in Figure 2, most algorithms did well in the case of detecting the existence and the width of aberrations for the large changes and high SNRs (upper left panels). For the cases of smaller aberrations and low SNRs, the smoothing methods (i.e. wavelets, lowess and quantile regression) gave better detection results (higher TPR and lower FPR) than other methods. The smoothing algorithms followed low amplitude and local trends in the data better than the other algorithms that were less sensitive to such features.

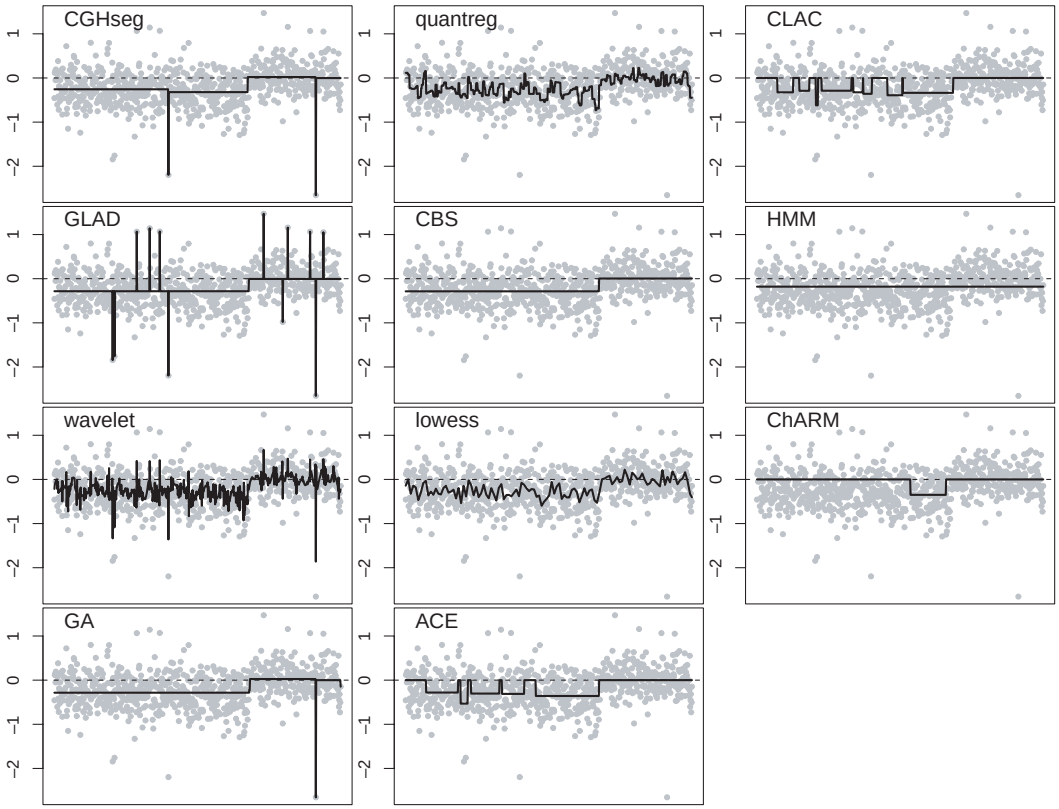

Fig. 3. Array-CGH profile of chromosome 13 in a Glioblastoma Multiforme sample (GBM31). This chromosome has a partial loss of low magnitude. Most algorithms in the study detect the loss. In particular, CGHseg, GLAD, CBS and GA clearly identify the region.

Among the methods that perform estimation (indicated by 'E' in Table 1), the homoscedastic algorithm in Picard *et al*. (2005) appeared to perform better than other methods. However, none of the algorithms reliably detected the aberrations with small width and low SNR because the signal is too weak to be differentiated from the noise. Compared with other algorithms, the CLAC algorithm tended to overestimate the boundaries of the aberrations. The mean smoothing step in CLAC appeared to reduce noise in the artificial chromosome data at the expense of blurring the edges of the boundaries. The ChARM algorithm did not detect the presence of aberrations in every artificial chromosome, even in the large width, high SNR case, irrespective of the cutoff p-values for the mean and sign tests. We suspect that this may be due to the fact that its boundary detection step is local and decoupled from the overall estimation.

Glioblastoma Multiforme (GBM) data

There are 26 samples representing primary GBMs in the glioma data from Bredel *et al*. (2005). GBM is a particularly malignant type of brain tumor, with a median patient survival time of a year. The samples were co-hybridized with pooled human controls onto custom spotted cDNA microarrays. The scanned raw data were downloaded from the Stanford Microarray Database (http://smd.stanford.edu). For the purpose of this paper, the array data were normalized by print-tip group, intensity-dependent normalization with the Limma package (Smyth, 2004). Of the 41 421 elements on each array, we were able to link 33 599 to chromosomal positions using mapped EST data from the hg16 build of the UCSC Genome Browser (http://genome.ucsc.edu). Missing values in each array were removed to avoid the effect of imputed values in subsequent analyses.

Though noisy (standard deviation of the log-ratios for each array ranges from 0.35 to 0.9) the GBM data contained a mixture of larger, low amplitude regions of gains/losses and smaller, high amplitude regions of amplifications/deletions. These types of copy number alterations represent the types of aberrations the array CGH algorithms should detect. Two examples representing a broad, low amplitude change and a smaller, high amplitude one are examined in the following paragraphs.

Numerous regions of gains/losses have been found in many microarray studies on gliomas (Koschny *et al*., 2002). For instance, gain of chromosome 7 and losses of chromosomes 10 and large portions of 13 and 22 have been observed in GBMs previously. These gains and losses may be the effect of uncontrolled mitotic events from point mutations of oncogenes and tumor-suppressing genes. In the sample GBM31 there exists a large region of loss on chromosome 13. The overall magnitude of the loss is very low because not all tumor cells in a given sample have the same types of gains and losses. It may also be due to the presence of connective tissues and other non-tumor cells in the sample. As a consequence of sample heterogeneity, the signal is diluted, thus complicating the detection procedure for the algorithms.

As can be seen in Figure 3, most algorithms in the study detected the proximal loss of chromosome 13 of GBM31. CGHseg, GLAD,

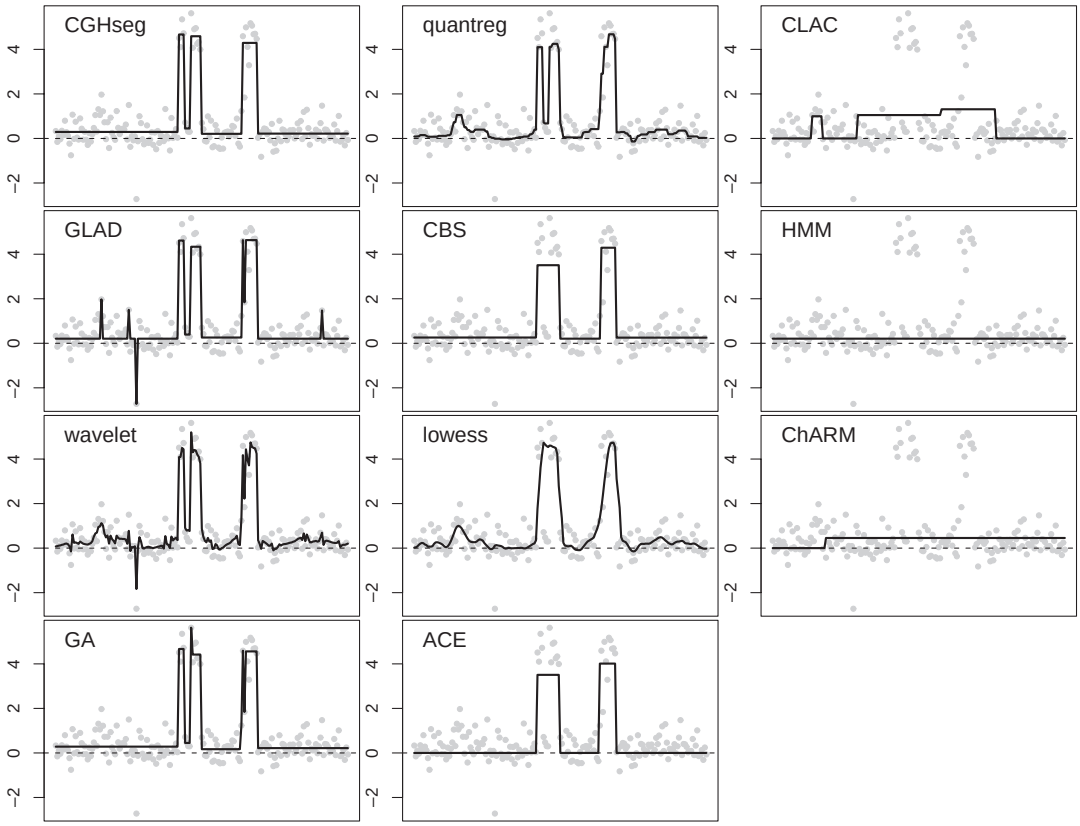

Fig. 4. Array-CGH profile of the three amplifications around EGFR in GBM29. CGHseg, quantreg, GLAD, wavelet and GA detect all three amplifications. CLAC, CBS, Lowess and ACE detect the first two amplifications as one larger region. ChARM detects the amplification as one large region of gain, while HMM does not detect any.

CBS and GA, all clearly identified nearly identical regions. All three smoothing algorithms showed the same general trend but the global loss was obscured by the local features. These algorithms performed well in detecting the smaller aberrations in the simulated data, but they were not as useful for a global view. HMM did not separate chromosome 13 into two regions. In addition to detecting the loss, GLAD identified numerous single-probe outliers. Such outliers can either indicate a real focal aberration, some type of polymorphism, or an experimental artifact (e.g. bad probe). CLAC and ACE detected the region of loss as a series of smaller losses. Many smaller regions within chromosome 13 that CLAC and ACE did not detect as losses coincided with localized positive spikes in log-ratios.

The GBM data also contained numerous amplifications. Several amplifications, such as those around PDGFRA, CDK4 and MDM2, have been well-studied in GBMs (Kraus *et al*., 2002). The amplification at the EGFR locus has been implicated in other tumors, and it is clearly present in GBMs, as shown in Figure 4. In this GBM29 sample, there appeared to be at least three high amplitude amplifications around EGFR. The algorithms CGHseg, quantreg, GLAD, wavelet and GA detected all three high amplifications. Because there are only four probes separating the first two amplifications, methods such as CLAC, CBS, Lowess and ACE combined the first two amplifications together. It is possible that these two amplifications were in fact a single one, but mapping the probes to their physical positions suggested that they are likely to be two separate aberrations.

CLAC, ACE, and ChARM, all use mean smoothing as an initial step to denoise the data. Mean smoothing increases SNR at the cost of blurring the edges of the boundaries. Because of the blurring, CLAC detected the amplifications as two larger adjacent amplifications. ACE does not merge the amplifications the way CLAC did, as it has an additional step to compensate for the blurring in identifying the boundaries. More sophisticated smoothing methods appear to perform better in general.

ChARM detected the three amplifications as one large region of gain. However, when each region was manually tested for significance in their software, all regions were marked as significant by their mean and sign tests. This indicated that the boundary detection part should be improved. HMM did not detect any of the three amplifications, even though it detected smaller regions in the simulated data. Singular matrices in the iterations were often the source of problems.

DISCUSSION

For processing a large number of high-resolution arrays, the speed of the algorithms becomes an issue. The simple smoothing algorithms such as lowess and wavelets are the fastest, while HMM and CBS are the slowest. The fast local algorithms are $\mathcal{O}(N)$, where N is

the number of probes along the chromosome, and slow ones, which require long-range information, are generally $\mathcal{O}(N^2)$. For the typical BAC arrays, speed is not a concern, but it can be a problem for high-density oligonucleotide arrays.

A particularly helpful feature for future implementations of some algorithms would be to estimate the statistical significance of the detected copy number changes and then rank them accordingly. Some current implementations simply return the processed profiles but do not 'call' the detected regions as significant or not. When the genome-wide profile is scanned for potentially new regions of interest, quantitative statistics about the aberrations are critical in order to decide which region to pursue for further examination. The false discovery rate appears to be a natural framework, but only two of the algorithms (CLAC and ACE) have incorporated this so far.

There is an inherent difficulty in comparisons of complicated algorithms. Each method has its own set of parameters that must be tuned properly, and this often requires a detailed understanding of the algorithm. It is therefore possible for the proponents of one algorithm to argue that their method did not perform adequately simply because the parameters were not set at an optimal level. This is especially true when the developers of different algorithms were interested in arrays of different resolutions and noise levels, or were motivated by different types of biological problems. For example, researchers working on well-characterized cancers might be more interested in focal aberrations, whereas those working on less characterized cancers might be interested in broad changes. To the extent that this is true, the results of our analysis here should be taken with caution, and it is incumbent upon the user to understand the characteristics of each method. On the other hand, if the algorithm is very sensitive to the changes in parameters or if the complexity of the algorithm does not allow the user to determine the correct parameters easily, it may be legitimately viewed as a weakness.

There are several ways in which these algorithms can be improved. First, further refining of the methodology especially in the preprocessing of the data would be beneficial. We have found that some segmentation methods, especially CGHseg (Picard *et al*., 2005) and CBS (Olshen *et al*., 2004), appear to perform consistently well. When the noise is high, smoothing methods appear to work well, although their output is more difficult to interpret. An optimal combination of the smoothing step and the segmentation step may result in improved performance. Second, one piece of information that was not considered in any of the methods discussed here is the physical distance of the probes along the genome; uniform spacing is assumed currently. If two probes indicating the same direction of change are very far apart, the probability that they refer to the same alteration should be lower than if they had been closer. It is not clear how much improvement could result by incorporating this information, but it can only help, if done correctly. Third, because multiple samples are usually analyzed at the same time, it is important to summarize the overall results in a clear fashion. Although those aberrations present in a small fraction of patients can be important, those occurring with a higher frequency are more likely candidates for research. Finally, user-friendly software with visualization tools and links to other databases would be helpful. Currently, these functionalities are present only for a small number of programs implementing the algorithms. An investigator is more likely to be interested in a region if a gene is present in the region, and even more so if it is an oncogene. Links to such information would be invaluable.

ACKNOWLEDGEMENTS

This work was supported by a grant from the National Institute of General Medical Sciences (P.J.P.) and the National Institute of Neurological Disorders and Stroke (M.D.J.).

Conflict of Interest: none declared.

REFERENCES

Autio,R. *et al*. (2003) CGH-Plotter: MATLAB toolbox for CGH-data analysis. *Bioinformatics*, **19**, 1714–1715.

Beheshti,B. *et al*. (2003) Chromosomal localization of DNA amplifications in neuroblastoma tumors using cDNA microarray comparative genomic hybridization. *Neoplasia*, **5**, 53–62.

Bredel,M. *et al*. (2005) High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Res.*, **65**, 4088–4096.

Brennan,C. *et al*. (2004) High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Res.*, **64**, 4744–4748.

Cleveland,W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, **74**, 829–836.

Daruwala,R.-S. *et al*. (2004) A versatile statistical analysis algorithm to detect genome copy number variation. *Proc. Natl Acad. Sci. USA*, **101**, 16292–16297.

Eilers,P.H.C. and de Menezes,R.X. (2005) Quantile smoothing of array CGH data. *Bioinformatics*, **21**, 1146–1153.

Fridlyand,J. *et al*. (2004) Hidden Markov models approach to the analysis of array CGH data. *J. Multivariate Anal.*, **90**, 132–153.

Hodgson,G. *et al*. (2001) Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nat. Genet.*, **29**, 459–464.

Hsu,L. *et al*. (2005) Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**, 211–226.

Hupe,P. *et al*. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Ishkanian,A.S. *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.*, **36**, 299–303.

Jong,K., Marchiori,E., van der Vaart,A., Ylstra,B., Meijer,G. and Weiss,M. (2003) Chromosomal breakpoint detection in human cancer. In *Lecture Notes in Computer Science*, Springer-Verlag, Berlin, Vol. 2611, pp. 54–65.

Jong,K. *et al*. (2004) Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, **20**, 3636–3637.

Koschny,R. *et al*. (2002) Comparative genomic hybridization in glioma: a meta-analysis of 509 cases. *Cancer Genet. Cytogenet.*, **135**, 147–159.

Kraus,J.A. *et al*. (2002) Molecular genetic analysis of the TP53, PTEN, CDKN2A, EGFR, CDK4 and MDM2 tumour-associated genes in supratentorial primitive neuroectodermal tumours and glioblastomas of childhood. *Neuropathol. Appl. Neurobiol.*, **28**, 325–333.

Lingjaerde,O.C. *et al*. (2005) CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics*, **21**, 821–822.

Lucito,R. *et al*. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.*, **13**, 2291–2305.

Myers,C.L. *et al*. (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, **20**, 3533–3543.

Olshen,A.B. and Venkatraman,E.S. (2002) Change-point analysis of array-based comparative genomic hybridization data. *American Statistical Association Proceedings of the Joint Statistical Meetings*, American Statistical Association, Alexandria, VA, pp. 2530–2535.

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

Picard,F. *et al*. (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, 27.

Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, **37**, Suppl 11–17.

Pinkel,D. *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Pollack,J.R. *et al*. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, **23**, 41–46.

Pollack,J.R. *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA*, **99**, 12963–12968.

Polzehl,J. and Spokoiny,V. (2000) Adaptive weights smoothing with applications to image restoration. *J. R. Statist. Soc. B.*, **62**, Part 2, 335–354.

R Development Core Team (2004) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sebat,J. *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.

Sen,A. and Srivastava,M. (1975) On tests for detecting a change in mean. *Ann. Stat.*, **3**, 98–108.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.

Snijders,A.M. *et al*. (2003) Shaping of tumor and drug-resistant genomes by instability and selection. *Oncogene*, **22**, 4370–4379.

Solinas-Toldo,S. *et al*. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.

Wang,P. *et al*. (2005) A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 45–58.